Prediction of Paroxysmal Atrial Fibrillation (PAF) Onset through Analysis of Inter-beat Intervals (IBI)

By

Charles Q. Du

Submitted to the Department of Electrical Engineering and Computer Science
In Partial Fulfillment of the Requirements for the Degrees of
Bachelor of Science in Electrical Engineering and Computer Science
and Master of Engineering in Electrical Engineering and Computer Science
at the
Massachusetts Institute of Technology
May 9, 2003
Copyright 2003 Charles Q. Du. All rights reserved.

The author hereby grants to M.I.T. permission to reproduce and distribute publicly paper and electronic copies of this thesis and to grant others the right to do so.

Author .................................................. Department of Electrical Engineering and Computer Science May 9, 2003

Certified by ........................................... .................................................................

Rosalind W. Picard
M.I.T. Thesis Advisor

Accepted by .................................................. Arthur C. Smith
Chairman, Department Committee on Graduate Theses
Prediction of Paroxysmal Atrial Fibrillation (PAF) Onset through Analysis of Inter-beat Intervals (IBI)

By
Charles Q. Du

Submitted to the Department of Electrical Engineering and Computer Science
In Partial Fulfillment of the Requirements for the Degrees of
Bachelor of Science in Electrical Engineering and Computer Science
and Master of Engineering in Electrical Engineering and Computer Science

Abstract

PAF is a type of progressive cardiac arrhythmia that poses severe health risks, sometimes leading to ventricular arrhythmia and post-operative mortality. Some of the difficulties with treating PAF include screening for patients with the disorder, detecting episode occurrences, and predicting occurrences. To address these issues, electrocardiogram (ECG) data from the PhysioNet Online Database was used to develop a technique to screen, detect, and predict the onset of PAF. Methodologies explored included Hidden Markov Modeling on inter-beat intervals, entropy, and heart-rate spectrograms. Initial testing indicates the technique to be discriminant between PAF and non-PAF (possibly other cardiac disorder) patients (89% sensitivity and 55% specificity). Even more promising is its ability to discriminate between PAF patients and healthy individuals (89% sensitivity and 81% specificity). Both results are from data not involved in training. The IBI-based algorithm could be incorporated into medical devices with the potential of contributing to new healthcare technology.

Acknowledgements

I wish to thank my thesis advisor, Prof. Rosalind Picard, for all her cheerful encouragement, tireless support, and helpful criticisms. Her patient and insightful feedback over the course of eight revisions helped me tremendously in shaping my endeavors in this fascinating and critical field of study.

Next on my thank you list would be Prof. Roger Mark, for being a taskmaster and mentor both. Thank you for your instruction in physiology and for keeping me in shape this term. I swear the heavy lifting I did TAing for 022 was the only real exercise I had time for. And to think, I was even paid to run around!

Many thanks also to Yuan Qi and Ashish Kapoor. Yuan for always being willing to share his brilliant insights and works. Ashish for always being available, taking time from his own busy schedule to offer helpful advice.

Thanks should also go to Dr George Moody and Wei Zong, for critical help in collecting data and interpreting results.

And of course, to Amy, who fed me and kept me warm when I pulled my late night work sessions.
To my parents, Miting and Jin.

Mom, Dad, I’ll start working on a PhD… as soon as I figure out what I want it in.

Promise!
# Contents

## 1.0 Introduction

1.0 Introduction .......................... 8  
1.1 Paroxysmal Atrial Fibrillation .................................................. 8  
1.2 Prior Works ........................................................................ 11  
1.3 Specific Relevant Techniques ................................................... 12  
1.4 Statement of Goals ................................................................ 15  

## 2.0 Theory

2.0 Theory ........................................ 16  
2.1 Electrophysiology ............................................................. 16  
  2.1.1 Electro-Cardiograms ......................................................... 16  
  2.1.2 Inter-beat Intervals .......................................................... 18  
  2.1.3 Spectrograms ................................................................ 18  
2.2 Pattern Recognition ............................................................. 20  
  2.2.1 Hidden Markov Modeling .................................................. 21  

## 3.0 Methods

3.0 Methods .......................................... 22  
3.1 Toolkit ................................................................................. 22  
3.2 Raw Data .............................................................................. 23  
3.3 Terminologies ........................................................................ 28  
3.4 Techniques ............................................................................ 28  
  3.4.1 Prematurity Weighting ........................................................ 29  
  3.4.2 Discrete HMMs on Entropy ............................................... 30  
  3.4.3 Gaussian-mixture HMMs on IBIs ....................................... 30  
  3.4.4 Gaussian-mixture HMMs on Spectrogram ......................... 31  
3.5 Training and Testing ............................................................... 31  
  3.5.1 Event 1 Screening ............................................................ 32  
  3.5.2 Event 2 Detection ............................................................ 35  
  3.5.3 Event 3 Prediction ............................................................ 36
4.0 Results

4.1 Prematurity Weighting ........................................ 39
4.2 Discrete HMMs on Entropy ........................................ 40
4.3 Gaussian Mixture HMMs on IBIIs .............................. 41
    4.3.1 Event 1 Screening ........................................ 42
    4.3.2 Event 2 Detection ........................................ 43
    4.3.3 Event 3 Prediction ........................................ 44
4.4 Gaussian Mixture HMMs on Spectrogram ....................... 44

5.0 Discussion

5.1 Event 1 Screening .................................................. 46
5.2 Event 2 Detection .................................................. 47
5.3 Event 3 Prediction .................................................. 48
5.4 Entropy Failure Speculation ..................................... 48

6.0 Conclusion

7.0 Recommended Future Works

Appendix A: Complete HMM Results

Appendix B: Spectrograms (with 20 frequency bins)

Appendix C: Spectrograms (with 50 frequency bins)

References
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exponential weighting of PBs</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>Demonstration of exponential weighting scheme</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Electrical anatomy of the human heart</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>Comparing spectrograms of a PAF patient, a non-PAF patient, and a healthy individual</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>Raw data used in analysis</td>
<td>23</td>
</tr>
<tr>
<td>6a</td>
<td>Analysis of type-P, N, and Q IBI series histogram</td>
<td>26</td>
</tr>
<tr>
<td>6b</td>
<td>Analysis of type-P1 &amp; P2 IBI histograms</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>Processing of data used in Event 1 screening</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>Processing of data used in Event 2 detection</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>Processing of data used in Event 3 prediction</td>
<td>37</td>
</tr>
<tr>
<td>10</td>
<td>Discretization of entropy series</td>
<td>41</td>
</tr>
</tbody>
</table>
List of Tables

Table 1 - Wei’s “last submission” Results ........................................ 14
Table 2 - Prematurity Weighting Test Ranges .................................... 29
Table 3 - Results Overview............................................................. 38
Table 4 - Prematurity Weighting Results ........................................ 40
Table 5 – Discrete-output HMM on Entropy Results ......................... 41
Table 6 – Gaussian-mix HMMs on IBIs, Event 1 ............................ 42
Table 7 – Gaussian-mix HMMs on IBIs, Event 2 ............................ 43
Table 8 - Gaussian-mix HMMs on IBIs, Event 3 ............................. 44
Table 9 - Gaussian-mix HMMs on Spectrograms, Event 1 (freq res 50)… 44
Table 10 - Gaussian-mix HMMs on Spectrograms, Event 1 (freq res 20) ...45
Table 11 - Gaussian-mix HMMs on Spectrograms, Event 2................. 45
Table 12 - Gaussian-mix HMMs on Spectrograms, Event 3.............. 45
1.0 Introduction

1.1 Paroxysmal Atrial Fibrillation

Cardiac arrhythmic disorders have been known for over a hundred years, and atrial fibrillation (AF) in particular has now been recognized as the most common of all arrhythmias. Though comprehensive statistics are not available, best estimates reveal that it is probably present in more than 1% of the population. AF is estimated to be present in 4-6% of all people above the age of 65, but is also present in young and middle aged individuals [Stefaneli et al., 2002]. AF is associated with a high mortality rate, largely due to stroke and congestive heart failure [Wyse et al., 2001]. It is found incidentally in about 25% of all stroke admissions and has been shown to lead to poor control of the ventricular rate. In coronary bypass patients, AF may represent risk for immediate post-operative mortality as well as prolonged hospitalization. Usually AF is associated with certain symptoms: chest discomfort, fatigue, dizziness, palpitations, dyspnea, and syncope [Savelieva et al., 2000]. Interestingly, a significant portion of patients diagnosed with the disorder suffer from no obvious symptoms, and are only diagnosed incidentally during physical examinations, pre-operative assessments, or population surveys. Even in patients that report symptomatic episodes, Holter recording and transtelephonic recording have demonstrated a rate of asymptomatic episodes that exceeds symptomatic episodes by more than twelve-fold. Given that asymptomatic episodes probably have the same health risk, a significant portion of the aging population could potentially benefit from early detection and treatment.

Paroxysmal (sudden attack) AF is a progressive of atrial fibrillation that could lead to permanent AF or other cardiovascular disorders. Cases of paroxysmal AF pose an
addition problem for detection, as episodes may be intense but short periods that could otherwise go unnoticed. To record these events as they occur, Holter devices are usually used to record 24-hour ECG data and then analyzed for signs of arrhythmic disorder. Treatment for these patients could include pharmacological suppression [Jais et al., 2000], high-frequency pacing (HFP), or RF ablation. The primary therapeutic goal is the restoration and maintenance of normal sinus rhythm (NSR), which leads to optimized cardiac functions [Luederitz et al., 2000]. Recent medical technology advances include the implantable atrial defibrillators, which apply synchronized shock therapy upon detection of atrial fibrillation.

In practice, detection of atrial fibrillation by atrial defibrillators is a well-known problem. Patients that undergo the implantation procedures are known to have atrial fibrillation, as determined through screening using Holter recording. Detection by implanted devices is also easier since the contacts used are in close contact with the cardiac tissue. Different defibrillators might use different detection procedures. One example is the detection system developed by Swerdlow et al [Swerdlow et al., 2000]. This system classified cardiac patterns as atrial arrhythmic by tracking P-waves (pattern matching ECG recording with idealized P-waves), and reported proper detection in 88% of 190 AT (atrial tachycardia) episodes, and 98% of 132 AF episodes. Adjustment of sensitivity to P-wave amplitude accounts for tradeoff between false positives from far-field R-waves and false negatives from missed P-waves. Note again that attainment of such high detection levels is due in large part to the lack of signal distortion that results from implantation. Surface ECGs (which are used in this particular study) demonstrate much greater variability in performance due to variability in noise inputs. Further
considerations in practical detection of AF include distinguishing between AF and ventricular fibrillations, which is the focus of dual-chamber detection algorithms.

Various research projects have focused on prevention of atrial fibrillation, including a new approach based upon the genetic basis of the disease and curing through gene therapy. As part of the Human Genome Project, researchers have identified genes involved in diseases such as cardiomyopathy, Long QT syndrome, and atrial fibrillation. The scrutiny in this area could go a long way towards explaining AF amongst the young and middle aged, where no cause could be found in most cases. There is also potential for helping members of around 100 families around the world which have been identified with a familial form of the disease [Brugada et al., 1999]. Unfortunately, gene therapy techniques have yet to be fully developed, and prior genetic knowledge of susceptibility towards AF does not aid in predicting when exactly episodes will occur.

On a more traditional track, researchers have also looked at AF prevention through permanent pacing or periodic pacing. The theory of this approach is that pacing may suppress triggers of AF such as atrial premature beats, or perhaps reduce atrial stretch which appears to predispose to AF. It has also been suggested that atrial pacing may increase the benefits of antiarrhythmic drugs. To date, only small scale studies have been done to indicate that atrial pacing may maintain NSR in patients with paroxysmal AF as the primary or sole arrhythmia. In such cases, interventions were applied after detection of abnormally high resting parasympathetic tone, or periods of relative bradycardia. The results have been mixed, with some patients remaining free from AF, others required additional anti-arrhythmic medication, and a few developed permanent AF after pacemaker implantation [Sopher et al., 2000].
The focus of current treatment has been post-onset intervention, with an emphasis on specificity so as to prevent false positive detection [Schwartz et al., 2000]. An additional area that deserves more attention is in the development of a reliable PAF prediction scheme. If an accurate and reliable system of heart attack prediction could be developed, it would be a major breakthrough in healthcare efficiency, allowing for interventions that prevent episodes that would otherwise result in significant tissue damage. A prediction scheme that is both sensitive and specific to AF would not only serve to initialize intervention, but also give a reliable measure of the intervention’s effectiveness.

1.2 Prior Works

Much of the prior work studied and cited in this study has been derived from PhysioNet. PhysioNet is a public service of the Research Resource for Complex Physiologic Signals, funded by the National Center for Research Resources of the National Institutes of Health. It offers free access via the web to large collections of recorded physiologic signals and related open-source software, and was a strong motivating factor in directing research in cardiology.

Physionet organizes the annual Computers in Cardiology competitions, with the goal to stimulate effort and advance the state of the art in facing a single clinically significant problem, and to foster both friendly competition and wide-ranging collaborations. In the 2001 CinC competition, the challenge was to develop a fully automated method to predict the onset of paroxysmal atrial fibrillation/flutter (PAF), based on the ECG prior to the event.
The top scorers in the 2001 challenge were announced during the 25 September plenary session of Computers in Cardiology in Rotterdam. The top score and the award in PAF screening was obtained by Günther Schreier and colleagues of the Austrian Research Centers Seibersdorf (Graz, Austria), with a predictive accuracy \([\text{TruePos} + \text{TrueNeg} / \text{all}]\) of 82%. In PAF prediction, the top score was obtained by Wei Zong and colleagues at the Harvard-MIT Division of Health Sciences and Technology (Cambridge, Massachusetts, USA), with a sensitivity \([\text{TruePos} / \text{TruePos} + \text{FalseNeg}]\) of 79%.

The effectiveness of the algorithm proposed in this paper would use the CinC challenge as one of its benchmarks, however, it is important to note the scoring metric used in the Prediction part. The score was not calculated based on the predictive accuracy as a whole, but instead on the number of consecutive ECG pairs from PAF patients that were correctly labeled as preceding or distal to an episode. The submitted labels for non-PAF ECGs were not used in scoring at all. Thus, the scoring on Prediction entirely focused on sensitivity, and was therefore biased toward schemes that considered each ECG pair to be from a PAF patient. More detailed description of the CinC test set will be covered in Section 3.2. Also missing from the data provided were continuation sets from the ECGs in the CinC test set, limiting the amount of data available for the purposes of PAF detection.

1.3 Specific Relevant Techniques

The technique used in the competition by Wei Zong and colleagues at MIT [Zong et al., 2001] was especially insightful. They noted that the number and timing of atrial premature beats (APBs), appeared to be "of significant value in terms of predicting
imminent PAF episodes”. In their analysis, they used an exponential weighting system that assigned a non-equal weighting value for each APB in an immediately preceding interval (Fig 1), with higher weighting given to APBs that occurred a short time prior.

As could be seen, the parameters used in determining the weighting and classification include apb-threshold, window length (w), and exponential rate (tau). The APB-threshold determines the ratio of an IBI over the time-averaged IBI length where the beat would be defined as an APB.

The time-averaged IBI length (RRavg) at time \( n \) is derived by the formula: 
\[
RRavg(n) = 0.9 \times RRavg(n-1) + 0.1 \times RR(n).
\]
For example, if the APB-threshold is 20%, and the RRavg at the time is 1 sec, any immediately following beats with an IBI shorter than 0.8 seconds would be designated an APB. Note: strictly speaking, this would indicate a premature beat, not necessarily than an atrial premature beat. Presumably, Wei’s beat detector ignored other types of premature beats (ventricular, junctional, etc.). For the purposes of this paper, these cardiac events will be referred to as Premature Beats, or PBs. The window length determines how far back in time to look for APBs. The exponential rate determines the relative weighting of APBs closer to the current time, as compared to those in the more distant past. Through a simple final PAF-threshold
analysis on the weighted results, they were able to produce a method of discrimination with high sensitivity and moderately high specificity.

Their utilization of weighted PB count was the original inspiration for this paper, which focuses on inter-beat intervals, a measure that is the time between heart beats, which is significantly smaller in the instance of an PB. With Wei's permission, his "last submission" answers were analyzed and are presented (Table 1). The results are presented in terms of specificity (true-negatives/ [true-negatives + false-positives]), sensitivity (true-positives/ [true-positives + false-negatives]), and predictive accuracy ([true-positives + true-negatives] / [all results]).

<table>
<thead>
<tr>
<th>Wei's Results</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens.</th>
<th>Spec.</th>
<th>Pred.Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening*</td>
<td>22</td>
<td>14</td>
<td>8</td>
<td>6</td>
<td>79%</td>
<td>63%</td>
<td>72%</td>
</tr>
<tr>
<td>Prediction</td>
<td>22</td>
<td>44</td>
<td>28</td>
<td>6</td>
<td>79%</td>
<td>61%</td>
<td>66%</td>
</tr>
</tbody>
</table>

Table 1 - Wei's “last submission” record, not his best result, which was 79% Predictive Accuracy

For the CinC competition, where the test data sets were pairs of ECGs, the algorithm described above was easily adapted. For Screening, the weighted value was computed for each ECG, and the maximum value between each pair was used. For Prediction, the weighted value for each ECG could be computed, and the ECG with the higher value is designated as pre-PAF.
In fact, Wei also noticed the bias towards sensitivity in the competition for prediction. He did not utilize any screening for non-PAF patients. Instead, he achieved a high score simply scoring each ECG pair based on the ECG with the higher APB count. This approach, unfortunately, would not work in real world scenarios.

1.4 Goals

This study has multiple goals. We wish to develop a system that does a good job with screening, detection, and prediction. To determine how well the system achieves each of these goals, we set up a series of test events to demonstrate its viability. Here, we will state the goals, and then outline the testing procedures in Section 3.3.

Test Event 1: PAF screening

Event 1 is to determine if subjects at risk of PAF can be distinguished from those representing a larger population, based on their ECGs. The test will involve comparing those ECGs in Group P (PAF), against those in Group N (non-PAF, with other disease), as well as those in Group Q (normal).

Test Event 2: PAF detection

Event 2 is to determine if subjects identified as at risk of PAF could be reliably identified as currently suffering a PAF attack or not. The test will involve those ECGs in Group P-Ic (no-current-attack PAF) against those in Group P-2c (current-attack PAF), as well as against those in Groups N-c and Q-c.
**Test Event 3: PAF prediction**

Event 3 is intended to determine if subjects in Group P have distinctive and detectable changes in their ECGs immediately before PAF. (In other words, is the imminent onset of PAF predictable in an individual known to be at risk of PAF?) A successful method for doing so should be able to determine for those in Group P-2c the detectable “precursors” for those about to suffer an attack.

**2.0 Theory**

2.1 Electrophysiology

The normal functioning of the cardiovascular system depends in large part on its electrical activity. The heart as a whole is controlled by the coordinated electrical activity of its cells. This control must be communicated between what are called pacemaker cells of the heart, both to determine the rate of beating, and to influence the strength of contraction. Disturbances of the pacemaker communications disrupt its normal ability to generate and conduct electrical signals. These disturbances give rise to abnormal electrical activity, which frequently leads to cardiac malfunction and death.

2.1.1 Electro-Cardiograms

To monitor the electrical behavior of the heart, cardiologist use a system called Electro-Cardiograms, or ECGs for short. It is at its most basic a system of electrodes used to measure electrical conduction on the skin. An ECG could be taken with as few as two “leads”, which are wires attached to the skin. By convention though, leads are
positioned on the body in such a way that six vertical and six horizontal “views” of electrical activity could be monitored and presented.

Figure 3
Electrical Anatomy of the Human Heart.
Note the correspondence between electrical stimulation and ECG readouts.

In general, different phases of the ECGs correspond to cardiac excitation in different areas of the heart. The initial structure observed for a given heartbeat is the P wave, which results from the spread of excitation through the atria from the SA node. During the P-R interval that follows, the atria contract to expel blood into the ventricles as the electrical impulse propagates down to the AV node. A sharp spike called the QRS complex is observed in the ECG at the end of atrial contraction, followed immediately by a Q-T interval during which ventricles contract. The last structure observed is normally the T wave, which results from the repolarization of the ventricles.

A full ECG data stream usually consists of several leads containing electrical levels sampled at around 128Hz. The resulting information is also subject to noise, both from
muscular distortions and problems with the wire attachment. Problems would then arise from analyzing the noisy data.

2.1.2 Inter-beat Intervals

The different analysis techniques presented in this study focus on inter-beat intervals (IBIs), which are the measured time intervals between QRS complexes. In physiological terms, IBIs correspond to the time between initiation of consecutive ventricular contractions. Part of the advantage of IBIs is also its relative detection reliability in the presence of various noise sources.

The physiological significance of Inter-beat Intervals has been established in studies of multifocal triggering and irregular nodal conduction. Interatrial septum pacing in particular [Padeletti et al., 2000] has also been demonstrated to be a safe and feasible technique for reduction of arrhythmia in general as well as associated mortality.

2.1.3 Spectrograms

Efforts were also made to study additional features of the data, such as the spectrogram and its entropy. Since the entropy is a scalar measure of the spectrum, a good spectrum estimation technique was needed. For the field of bio-medical applications, the signals are non-stationary, and therefore require non-standard spectrum estimation techniques. To this end, Yuan Qi’s algorithm [Qi et. al., 2002] was used, with results to be mentioned in Section 4.2.

The spectrogram could be thought of as an application of Short-Time Fourier Transforms, with the frequencies on the y-axis and the time segments on the x-axis. The
power in each frequency at any particular time instant is represented by the amplitudes of 
the vertical cross-sections at that time instant. In standard literature, the power spectrum 
of heart rate for a normal person could be found represented by three frequency bands. 
The low-frequency, medium frequency, and high-frequency bands range from 0 to 0.5 
Hz, and have been linked to various sympathetic and parasympathetic control 
mechanisms. Due to the complex nature of biological controls and the natural variability 
between individuals and physiological states, it is yet unclear how competing influences 
could be segregated. In the case of patho-electrocardiology in particular, it is unclear 
how the frequencies of activity are influenced by the myriad of ectopic activities 
associated with arrhythmias, each of which would introduce variability to instantaneous 
heart rate (ihr) measurements. For example, the parasympathetic controls might detect the 
sudden acceleration in SA nodal activity and act to suppress heart rate as a whole. 
Another possibility would be the atrial irregularities resulting in insufficient atrial filling, 
triggering the baro-receptor reflex to increase heart rate to compensate. It is hard to 
predict which effect would dominate.

On the spectrogram, heart rate irregularity due to any input would appear as 
contributions to all frequency bands, or vertical streaks (Figure 4). In order to take 
advantage of the spectrogram features, the tradeoff would need to be made between time 
resolution and frequency resolution. Time resolution would be favored to better 
determine the duration of ectopic behavior. On the other hand, to better examine the 
changes in power spectrum preceding and following ectopic events, greater frequency 
resolution would be necessary regardless of actual control influences (sympathetic, 
parasympathetic, etc).
Entropy (defined as $E[-\log p(x)]$) is a relative measure of variability present in the power spectrum in the sense that it is maximal for a uniform distribution. In the case of ectopic beats, one would expect high entropy values corresponding to times of episodic occurrences.

2.2 Pattern Recognition

Pattern recognition is a field related to signal processing and artificial intelligence. It is the general name given to a diverse set of techniques with which one could classify and model signals. In general, there are two types of signal models which are utilized, known as deterministic models and statistical models. Deterministic models generally exploit some known properties of the signal, whereas statistical models try to characterize the signal as a parametric random process.
2.2.1 Hidden Markov Models

Hidden Markov Models are a particular form of statistical models. In this specific application, the HMM will be used to model the observations (features of time series of IBIs) corresponding to the hidden physiological states (degree of PAF-behavior). Compared with observable Markov models, HMMs have the advantage of being widely applicable and could be applied with few presumptions about the underlying hidden states. Every HMM has a set of parameters, which could be trained to better account for certain sets of observed behavior. In the toolkit by Kevin Murphy (Section 3.1), the main parameters are $Q$ (number of states in the model), $M$ (number of Gaussian mixtures), $O$ (number of possible output observations), and "diagonal" or "full" (for covariance matrices). Other important aspects of setting up the problem, such as initialization of states, state transition matrices, and observation probability distribution in states are handled by the toolbox. Given a set of initial state parameters, an HMM could be "trained" on sets of observation sequences to better describe the evolution of observations (the number of times the HMM loops over each piece of training data is yet another parameter). After training HMMs on different sets of observations, we could then evaluate different observation sequences using the HMMs. Each HMM would yield a probability of that particular observation sequence arose from its model, and the HMM that returns the highest probability is the one that best fits the observations. Two particular types of HMMs used in this study are the discrete-output HMMs and Gaussian-mixture HMMs.
Discrete-output HMMs model events with discrete observations. The range of observations fit a finite “alphabet”, with the drawback that quantization of observations could lead to loss of useful information. Gaussian-mixture HMMs are applied towards modeling of outputs better characterized as samples out of a continuous distribution. By adjusting the number of states and the number of Gaussian-mixtures, the model could account for a wide range of observations. One of the parameters on the Gaussian-mixture HMMs is “left-right”, which allows for modeling using a particular form of state transitions. If set to “left-right”, then the context would be a series of states where transition is forbidden to any state with lower indices. In terms of physiology, this could be used to characterize progressive changes over time.

3.0 Methods

Throughout the period of this research project, several public domain toolkits were utilized to aid in the analysis of data. Each has proven to be quite useful, and will be briefly presented in Section 3.1. The availability of these tools significantly helped in speeding up the analysis. Description of the raw data are presented in section 3.2. The way the raw data was processed and organized, as well as how techniques were applied to training and testing are presented in Section 3.3.

3.1 Toolkits

WAVE is an extensible interactive graphical environment for manipulating sets of digitized signals with optional annotations. Designed for workstations with the open
source Xview toolkit, WAVE was built using the WFDB library developed for physiologic signal processing, so it could be applied to any of a wide variety of data formats supported by the WFDB library. In addition, a beta version of WAVE re-implemented using the Gimp Tool Kit (GTK), which was portable to MS Windows, was also used. The current release of GTKWave is still in development and should be considered usable but potentially unstable. WAVE was used in pre-processing the ECGs to extract the IBI information.

In addition to WAVE, the Hidden Markov Model (HMM) Toolbox for Matlab written and provided by Kevin Murphy was also used extensively. The functions were easy-to-use m-files that allowed for models ranging from discrete-output HMMs to Gaussian-mixture HMMs. Both modeling techniques were explored in the course of the analysis.

3.2 Raw Data

Figure 5 - Raw data. note that all PAF episodes (if any) were only present in continuation sets.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>P-cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group P</td>
<td>30-minute ECGs, PAF, 25 sets type P, 25 sets type P-2</td>
<td>5-minutes, 50 sets</td>
</tr>
<tr>
<td>Group N</td>
<td>30-minute ECGs, non-PAF, possibly other cardiac disorder, 50 sets</td>
<td>5-minutes, 50 sets</td>
</tr>
<tr>
<td>Group Q</td>
<td>24-hour ECGs, considered non-antithetic, 18 sets</td>
<td></td>
</tr>
</tbody>
</table>

The entirety of the data consists of 4 Groups (Figure 5), organized according to the types of patients. Group P consists of data from patients with PAF. Group N contains data from patients with no PAF, but perhaps some other cardiovascular disease. Group Q consists of data from the MIT-BIH
Normal Sinus Rhythm Database, recorded from patients with no detectable arrhythmias. Finally, a test set that was originally used for the CinC challenge was incorporated to study the relative effectiveness of the technique developed in this study. Each record contains 30-minutes worth of ECG recordings. Each of the records in Group P and Group N has an additional corresponding 5-minute "continuation" record. A more detailed description is as follows.

The Group P records contains 25 type P-1 records and 25 type P-2 records. P-2 records contain the ECG immediately preceding an episode of PAF, which can be verified by examining the like-numbered continuation record (type P-2c). Thus, for example, record p16 (a type P-2 record) immediately precedes the episode of PAF in record p16c. The records of type P-1 contain 30 minutes of the ECG during a period that is distant from any episode of PAF (there is no PAF episodes during the 45-minute period before the beginning or after the end of the 30-minute record). The corresponding 5-minute continuation record (type P-1c) shows that (at least!) the minutes immediately following the record do not contain a PAF episode. Since the data was collected with no manual audits, a few of the 30-minute records in this group may contain very short bursts of PAF episodes that escaped notice while the learning set was being compiled.

The 50 records in Group N come from subjects who do not have documented atrial fibrillation, either during the period from which the records were excerpted or at any other time. The subjects include healthy controls, patients referred for long-term ambulatory ECG monitoring, and patients in intensive care units with possibly other cardiac problems. It was interesting to note that during preliminary analysis, the author
of this report printed IBI histograms of the presumed “normal” population only to
discover obvious signs of arrhythmia. (Figure 6)

Since the absence of PAF alone does not establish “normal” electrical
cardiovascular behavior, additional data not included in the original CinC challenge had
to be found from which benchmarks of strictly normal patterns could be inferred.

Due to the discovery of abnormal ECG activity in the Group N patients, further
data was sought out online, also from the Physionet online database. The data from the
MIT-BIH Normal Sinus Rhythm Database were recorded from patients with no
detectable arrhythmias. These 18 ECGs are classified in this report as belonging to
Group Q, and used as a benchmark for comparison against the presence of abnormalities.
Each of the original records in this group represented 20-hours worth of ECG recording.
For the purposes of this study, four sections of recordings were taken from each 20-hour
record, and analyzed. The procedure with which the sections were extracted are
discussed in Section 3.3.
Figure 6a – Type-P (example is actually P-2) IBI series histogram (top), Type-N (middle), Type-Q (bottom), each across seven consecutive 200-IBI segments. The interval lengths are reflected by the x-axis, and the y-axis reflects the number of IBIs of that length. Note the pattern shows a tri-modal distribution suggestive of consistently irregular heart beats for type-N as well as type-P. A strictly normal IBI distribution would have a histogram consistently similar to that of type-Q, with a single Gaussian distribution and no outliers.
Figure 6b—Closer inspection of type P1 IBI histogram and its continuation P1c IBI histogram, as compared against that of type P2 and P2c. Note that the type P2 histogram contained greater spread, indicative of abnormal beats. Also, there was a noticeable distribution shift between the last 200-IBI histogram of P2 and P2c, whereas the distribution was consistent between P1 and P1c.
The final addition to the collection of data utilized was the test set provided by the original CinC challenge. This test set consisted of 100 30-minute records, which were classified either type P-1, P-2, or N. In the original competition, the test set was used to evaluate relative success in screening and prediction. It is used for the same purpose in this study. Since no continuation sets are available, no conclusions about detection could be drawn from this data set. Note that the ECGs for the CinC test set (only!) were grouped so that consecutive pairs belonged to a single patient. If the patient was a PAF patient, then one of the ECGs would be distal to any episodes while the other immediately precedes an episode.

3.3 Terminologies

In evaluating the results from the different tests, a set of common measures will be used to determine relative success. From the field of health studies, the terms specificity (true-negatives/ [true-negatives + false-positives]) and sensitivity (true-positives/ [true-positives + false-negatives]) will be used, as well as predictive accuracy or accuracy ([true-positives + true-negatives] / [all results]). From signal processing, the terms detection rate (synonymous with sensitivity) and false-alarm rate (1 – specificity) could also be used to characterize test results.

3.4 Analysis Techniques

Four different analysis techniques were used in exploring how best to discriminate between subsets of patients in each of the test event categories.
3.4.1 Prematurity Weighting

The first technique tried was a variant of the exponential weighting of PBs used by Wei Zong. The differences in the analysis were that 1) the IBI extraction in this study was done using WAVE, instead of the specially developed beat type detector used by Wei, 2) the definition threshold for an PB in this study was not fixed at 15% (as Wei assumed in his analysis), but was allowed to be a test variable. The parameters involved were: Tau, the exponential rate used in the weighting function; W, the length of time the exponential weighting function extends into the past; PB_threshold, the ratio of an IBI over the time-averaged IBI length where the beat would be defined as an PB; and PAF_threshold, the cutoff value of the weighted results between those designated as PAF or not. To determine the optimum parameters, the variables were tested over the following ranges (Table 2) in the training set. The optimum parameters determined were subsequently applied to the evaluation and testing sets.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial Value</th>
<th>End Value</th>
<th>Search Increment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tau</td>
<td>1</td>
<td>10</td>
<td>1/3</td>
</tr>
<tr>
<td>W</td>
<td>5</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>PB_threshold</td>
<td>.75</td>
<td>.95</td>
<td>.01</td>
</tr>
<tr>
<td>PAF_threshold</td>
<td>20</td>
<td>320</td>
<td>1</td>
</tr>
</tbody>
</table>

For the CinC test sets, the screening methodology followed Wei's original approach (Section 1.2), where the max weighted value of each ECG pair were taken to be representative of the person's state.
3.4.2 Discrete-output HMMs on Entropy

The second technique studied was the application of discrete-output HMMs on the entropy of the signal. The relevant parameters included both the number of discretized levels, and the number of states. The level of discretization was chosen to be 100, which resulted in as little information loss as possible (Fig 10). The number of states was varied between 2 and 15. For the generation of the entropy signal, the inputs were taken as the instantaneous heart rate, the corresponding relative time markers, and a frequency range for the spectrogram. The frequency range was varied between 2, 4, and 6 Hz with no observable changes in analysis results.

3.4.3 Gaussian-mixture HMMs on lBls

The third analysis technique was the one given the most in-depth exploration. Gaussian-output HMMs were trained on classification types and tested in each of the event categories. The relevant parameters were the number of mixtures fitted (M), the number of states in the model (Q), the type of transition matrix (left_right/not). The maximum number of iterations in training was set to be 5 for the entirety of the study, and the type of covariance matrix was set to ‘diag’. The parameters were varied (M between 1&5, Q between 1&3, and left_right between true/false), with all results that converged recorded in Appendix A. Important results were highlighted in the Section 4.3.
3.4.4 Gaussian-mixture HMMs on Spectrograms

Finally, a last analysis technique was attempted, also using Gaussian-output HMMs. Instead of IBIs, the models were instead trained and tested on spectrograms of the heart rate signal. The same parameters applies from the above subsection, except that the type of covariance matrix was set to be 'full'. Using covariance matrix type ‘diag’ resulted in numerical errors. Also of concern was the need to discard startup noise from the spectrogram estimation. To accomplish this, the algorithm was run with the full instantaneous heart rate input for the 30-minute ECG and the 5-minute continuation ECG appended together. The last 1400 samples of each output series were taken to be noise-free.

3.5 Training and Testing

For each of the tests, the raw data is processed (as will be detailed in the following sub-sections), and divided into training, evaluation, and testing sets. Except for the CinC Challenge test sets, which provides the metric of comparison, all other data series are randomly divided into the three mentioned dataset types (training, evaluation, and testing) with a probability ratio of 1:1:1 (so each data series has an equal probability of being assigned to any one of the 3 groups). The three groups (which would on average contain roughly the same number of datasets) could then be used to train, evaluate, and test the algorithm, respectively. Since it is not known if any one person contributed more than one data set, we cannot say with certainty that the people in the training set were different than those in the evaluation and test sets. In the case of HMMs, the optimum parameters are experimentally determined using the training and evaluation sets. The
algorithm would be applied to the test set and the CinC set only once using the determined optimum parameters. For the Prematurity Weighting method, the optimum parameters were determined on the training sets, and tested exactly once each on the evaluation and testing sets. This structure minimizes bias in test results.

3.5.1 Event 1: Screening

The training and testing was arranged differently for each Event. Event 1 (PAF screening) in particular was based on comparing results aimed at discriminating between Groups P, N, and Q, rather than focusing on Group P as in Events 2 & 3 (PAF detection & prediction, respectively).

The raw ECG data corresponding to the various patient groups were compiled and processed using the ihr, or instantaneous heart rate, function provided by WAVE. The resulting data is then inverted to derive 30 minutes worth of inter-beat intervals for each record. All data are then loaded into Matlab for processing and analysis.

Figure 7 – Event 1 screening. Each data series now contain uniform numbers of IBIs.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group P</td>
<td>1200 IBIs, PAF, 25 sets type P-1, 25 sets type P-2</td>
</tr>
<tr>
<td>Group N</td>
<td>1200 IBIs, non-PAF, possibly other cardiac diseases, 50 sets</td>
</tr>
<tr>
<td>Group Q</td>
<td>1200 IBIs, certified as non-arrhythmic, 72 sets</td>
</tr>
<tr>
<td>CinC test set</td>
<td>1200 IBIs, mixture of type P and type N ECGs, 100 sets</td>
</tr>
</tbody>
</table>

Further pre-processing is done by partitioning each record into 1200 IBI segments (with extra IBIs dropped from the beginning of the 30 minutes). Since each record reflects a patient with variable heart rate, every fixed-time interval segment would likely
contain a different number of IBIs. By establishing a fixed number of IBIs per segment, this discrepancy between segments could be eliminated, while still capturing the sequence of heart beats that led up to a possible cardiac episode. The IBIs discarded from the beginning of the time series are known to be distal to PAF episodes, and assumed to be less relevant to PAF observation. The segment length (in other words, the number of IBIs included), was chosen to be 1200 due to the fact that the number of IBIs contained in the different 30-minute time series of Group P & N ranged from 3379 to 1331. In other words, the time series with the slowest average heart rate contained 1331 detected beats, and 1200 is an arbitrarily chosen segment length that all ECG time series could satisfy.

The single exception to the 1200 IBI rule was one of the CinC test set series, which contained only 1142 IBIs over 30-minutes. The discrepancy is minor, and individual performance tracking of that specific data series did not indicate strong biases in results.

The training and testing across techniques are kept consistent to allow comparison. Prematurity weighting was the first technique attempted. It serves as the benchmark for later comparison with other methods. The algorithm was partially reproduced from the original paper (as cited in Section 1.3) and tested on the CinC test dataset, so as to verify the results originally reported.

The technique using discrete-output HMMs on entropy was implemented using the following procedure. Two separate HMMs were trained on the entropy estimates for Groups P & N (Fig 10), and then tested on the CinC test set. The data series used in this part of the analysis were derived from the full 30-minute ECGs using Yuan Qi’s
algorithm. The analysis was repeated using the 1200 IBI segments with no noticeable difference in results. Given the surprising lack of obvious features and results, the technique was only applied with Event I Screening. Attempts were made to probe into the reason behind the failure.

The technique using Gaussian-mixture HMMs on IBIs was more involved, with data from application to all 3 Events, and the promising results that were derived prompted further exploration. Group P and N data, as mentioned in Section 3.2, were each organized as 50 records, each with a segment of 1200 IBIs preceding PAF derived from the original 30-minute records. Group Q, which was originally eighteen 20-hour ECG recordings, was converted into IBI time series using the ihr function mentioned earlier. The first 8000 IBIs in each type-Q data series were separated into 4 consecutive sets of eighteen 2000 IBI segments. Each 2000 IBI segments were turned into 1200 IBI segments by truncating the last 800 IBIs of each segment. The end result was the 72 sets of 1200 IBI segments, as shown in Figure 7 above. Over the course of the IBI analysis, the maximum iteration parameter was kept at 5, and the covariance matrix type set to be diagonal.

Finally, the Event I testing was attempted using Gaussian mixture HMMs trained on spectrograms directly. This approach builds on the use of entropy as a discriminatory measure, looking beyond variability in the power spectrum, and modeling instead the appearance of ectopic beats and possible shifts in rhythmic control frequencies. For this part, the covariance matrix type was set to be full, since the datasets modeled are larger, with 20 frequency bins per time sample for the purposes of this study (so each set of
input observations were a 20x1200 array). Diagonal covariance matrix was attempted but generated too many numerical errors for the results to be useful.

3.5.2 Event 2: Detection

For Event 2 (PAF detection), processing similar to that described for Event 1 was applied to each of the 5-minute continuation sets, instead of the 30-minute sets (Fig 8). The resulting IBIs are collected into a single 200 IBI segment each (with extra IBIs dropped from the end of the 5 minutes). The IBIs at the beginning are of more interest, since the data specification as introduced in the CinC competition stated that the PAF episode (when present) would be at the beginning of the 5-minute segment that has been marked off. A visual presentation of processed data is presented below in Figure 8. Since the CinC test sets did not include continuation sets, they could not be used in testing for Event 2.

Note that one of the continuation sets of type N contained only 60 IBIs over 5 minutes. Checking the original ECG showed that the record flat-lined about a minute into recording, possibly due to lead detachment.
For training and testing, only Gaussian-mixtures were used. Separate HMMs were trained on the type P-1c, P-2c, N, and Q sets, and then evaluated for log-likelihood matching against the 200 IBI segments in the evaluation and test sets. Note the distinction emphasized here between type P-1c’s (no PAF episode) and type P-2c’s (PAF episode). The ultimate goal is to distinguish type P-2c ECGs from all others. To accomplish this, the algorithm for randomized set assignments and testing was actually run twice. The first time testing for pair wise discrimination between P-2c, N, and Q, and the second time for pair wise discrimination between P-2c, P-1c, and Q. (Fig 8)

3.5.3 Event 3: Prediction

Event 3 (PAF prediction) used similar pre-processing procedures on the raw data as the previous test events. The time series were converted into IBIs and segmented. In this test case, however, each of the 1200 IBI segments were further divided into six consecutive 200 IBI segments (Fig 9).

The goal of the exercise was to determine whether a patient would soon be experiencing a PAF episode, before the episode occurs. In terms of this study, that means singling out the type P-2 datasets. The HMMs were trained on the first and last 200 IBI
segment (assumption is that the last 200 IBI segment prior to PAF episode would contain features that could be used for prediction) of each data series. The HMMs are then used to evaluate the log likelihood of matching the segments in the middle. If the segments in the middle were consistently more similar to the first segment (a higher log likelihood), then the record was classified as imminently-PAF. After trying out different comparative methods (using different combinations of segments), it was found that the 200 IBI segment immediately prior to the last 200 IBI segment was the most discriminant (which is what would be expected). This procedure was noteworthy in that the attempt was made to train on acceptable state and episodic states for an individual, rather than a population of patients.

Figure 9 - Event 3 prediction. The long 1200-IBI segments were truncated and the resulting 200-IBI segments are used for predictive testing.

<table>
<thead>
<tr>
<th>Group P</th>
<th>six 200-IBI segments, PAF, 25 sets type P-1, 25 sets type P-2</th>
<th>P-cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>200 IBIs, 50 sets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group N</th>
<th>six 200-IBI segments, non-PAF, possibly other cardiac disorders, 50 sets</th>
<th>N-cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>200 IBIs, 50 sets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group Q</th>
<th>six 200-IBI segments, certified as non-arhythmic, 72 sets</th>
<th>Q-cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>200 IBIs, 50 sets</td>
</tr>
</tbody>
</table>

The results from using the segmented tracking were found to be useful, but yet another classification was explored using the general form used in Event 1 Screening. Separate HMMs were trained on P-2 types (the patients about to suffer a PAF episode).
P-I types (PAF patients not about to suffer an episode), and Q types. Training was done using the contiguous 1200 IBI segments used previously in Event 1.

4.0 Results

The complete results of all parameters and techniques explored are presented in Appendix A. The important results are presented in order of techniques used in the sections to follow. Here is a quick overview of the best results achieved across the different technique used for the set of test events:

Table 3 - Results Overview  * discontinued due to poor results  ** result cited in Abstract  ***results from CinC test set (see Section 4.3.1) yielded even better results, and were cited in Abstract

<table>
<thead>
<tr>
<th></th>
<th>Prematurity Weighting</th>
<th>Discrete-output HMM(Entropy)</th>
<th>Gaussian-mix HMMs(IBIs)</th>
<th>Gaussian-mix HMMs (Spectrograms)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Screening</strong></td>
<td>Sensitivity:94</td>
<td>Sensitivity:75</td>
<td>Sensitivity:88***</td>
<td>Sensitivity:87</td>
</tr>
<tr>
<td></td>
<td>Specificity:64</td>
<td>Specificity:14</td>
<td>Specificity:54</td>
<td>Specificity:24</td>
</tr>
<tr>
<td></td>
<td>Accuracy: 81</td>
<td>Accuracy: 48</td>
<td>Accuracy: 72</td>
<td>Accuracy: 53</td>
</tr>
<tr>
<td><strong>Detection</strong> (P2c vs. Nc)</td>
<td>Sensitivity:100</td>
<td>*</td>
<td>Sensitivity:100</td>
<td>Sensitivity:13</td>
</tr>
<tr>
<td></td>
<td>Specificity:93</td>
<td>Specificity:94</td>
<td>Specificity:79</td>
<td>Specificity:89</td>
</tr>
<tr>
<td></td>
<td>Accuracy: 95</td>
<td>Accuracy: 85</td>
<td>Accuracy: 88</td>
<td>Accuracy: 67</td>
</tr>
<tr>
<td></td>
<td>Specificity:85</td>
<td>Specificity:94</td>
<td>Specificity:67</td>
<td>Accuracy: 50</td>
</tr>
<tr>
<td></td>
<td>Accuracy: 87</td>
<td>Accuracy: 87</td>
<td>Accuracy: 87</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Prematurity Weighting</th>
<th>Discrete-output HMM(Entropy)</th>
<th>Gaussian-mix HMMs(IBIs)</th>
<th>Gaussian-mix HMMs (Spectrograms)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Screening</strong></td>
<td>Sensitivity:93</td>
<td>*</td>
<td>Sensitivity:89</td>
<td>Sensitivity:93</td>
</tr>
<tr>
<td></td>
<td>Specificity:44</td>
<td>Specificity:55</td>
<td>Specificity:27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy: 67</td>
<td>Accuracy: 55</td>
<td></td>
<td>Accuracy: 59</td>
</tr>
<tr>
<td><strong>Detection</strong> (P2c vs. Nc)</td>
<td>Sensitivity:91</td>
<td>*</td>
<td>Sensitivity:100</td>
<td>Sensitivity:9</td>
</tr>
<tr>
<td></td>
<td>Specificity:82</td>
<td>Specificity:65</td>
<td>Specificity:94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy: 86</td>
<td>Accuracy: 79</td>
<td>Accuracy: 61</td>
<td></td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>Sensitivity:71</td>
<td>Sensitivity:42</td>
<td>Sensitivity:43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specificity:56</td>
<td>Specificity:100</td>
<td>Specificity:75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy: 62</td>
<td>Accuracy: 74</td>
<td>Accuracy: 67</td>
<td></td>
</tr>
</tbody>
</table>
### 4.1 Prematurity Weighting

For this part of the analysis, the technique utilized was intended to reproduce and further expand upon the technique described by Wei Zong’s paper (see Section 1.2). The end results of the simulation (at 76% classification accuracy ([True Pos + True Neg] / Total) for Event 1) closely matched the results (79% classification accuracy) reported by Wei (though the “last submission” records on Physionet only showed a 72% classification accuracy, see Wei’s Results in Section 1.3). The end parameters were a threshold of 317, with a PB defined as an IBI that is 23% shorter in duration than the weighted sum of previous IBIs. The difference in beat detectors used was probably the

<table>
<thead>
<tr>
<th>P vs. Q</th>
<th>Prematurity Weighting</th>
<th>Gaussian-mix HMMs(IBIs)</th>
<th>Gaussian-mix HMMs(Spectrograms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening</td>
<td>Sensitivity:94</td>
<td>Sensitivity:88 **</td>
<td>Sensitivity:57</td>
</tr>
<tr>
<td></td>
<td>Specificity:93</td>
<td>Specificity:92</td>
<td>Specificity:93</td>
</tr>
<tr>
<td></td>
<td>Accuracy:93</td>
<td>Accuracy:90</td>
<td>Accuracy:81</td>
</tr>
<tr>
<td>Detection</td>
<td>Sensitivity:100</td>
<td>Sensitivity:100</td>
<td>Sensitivity:15</td>
</tr>
<tr>
<td>(P2c vs. Qc)</td>
<td>Specificity:85</td>
<td>Specificity:91</td>
<td>Specificity:100</td>
</tr>
<tr>
<td></td>
<td>Accuracy:87</td>
<td>Accuracy:93</td>
<td>Accuracy:74</td>
</tr>
<tr>
<td></td>
<td>Specificity:95</td>
<td>Specificity:90</td>
<td>Specificity:77</td>
</tr>
<tr>
<td></td>
<td>Accuracy:94</td>
<td>Accuracy:84</td>
<td>Accuracy:70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P vs. Q</th>
<th>Prematurity Weighting</th>
<th>Gaussian-mix HMMs(IBIs)</th>
<th>Gaussian-mix HMMs(Spectrograms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening</td>
<td>Sensitivity:95</td>
<td>Sensitivity:89</td>
<td>Sensitivity:53</td>
</tr>
<tr>
<td></td>
<td>Specificity:91</td>
<td>Specificity:81</td>
<td>Specificity:95</td>
</tr>
<tr>
<td></td>
<td>Accuracy:93</td>
<td>Accuracy:84</td>
<td>Accuracy:78</td>
</tr>
<tr>
<td>Detection</td>
<td>Sensitivity:100</td>
<td>Sensitivity:100</td>
<td>Sensitivity:6</td>
</tr>
<tr>
<td>(P2c vs. Qc)</td>
<td>Specificity:91</td>
<td>Specificity:78</td>
<td>Specificity:100</td>
</tr>
<tr>
<td></td>
<td>Accuracy:94</td>
<td>Accuracy:83</td>
<td>Accuracy:44</td>
</tr>
<tr>
<td>Prediction</td>
<td>Sensitivity:80</td>
<td>Sensitivity:82</td>
<td>Sensitivity:30</td>
</tr>
<tr>
<td></td>
<td>Specificity:96</td>
<td>Specificity:87</td>
<td>Specificity:83</td>
</tr>
<tr>
<td></td>
<td>Accuracy:94</td>
<td>Accuracy:85</td>
<td>Accuracy:64</td>
</tr>
</tbody>
</table>
main variability factor, and might have resulted in the slight (3\%) reduction in classification accuracy.

In comparing these results to the results to follow, it is useful to note that Wei Zong’s APB method and the modified Prematurity Weighting method used here only utilizes the last 10 minutes of the 30-minute ECG data. In later analysis using HMMs, when some IBIs are dropped from the beginning of the 30-minute ECG, the results of the PB method still serve as a useful benchmark. The following table presents the results tested on the CinC data set using optimized parameters derived from training sets.

**Table 4**

<table>
<thead>
<tr>
<th>Event 1: Screening</th>
<th>CinC Set</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>PredAcc</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>Full 30-min segments</td>
<td>19</td>
<td>19</td>
<td>3</td>
<td>9</td>
<td>68%</td>
<td>86%</td>
<td>76%</td>
<td>paf_threshold=317 pb_threshold=.77 window = 10 tau = 6.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event 2: Detection</th>
<th>CinC Set</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>PredAcc</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. P1c</td>
<td>Full 5-min segments</td>
<td>19</td>
<td>22</td>
<td>3</td>
<td>6</td>
<td>76%</td>
<td>88%</td>
<td>82%</td>
<td>paf_threshold=24 pb_threshold=.75 window=5 tau=6.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event 3: Prediction</th>
<th>CinC Set</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>PredAcc</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>Full 30-min segments</td>
<td>17</td>
<td>53</td>
<td>19</td>
<td>11</td>
<td>61%</td>
<td>74%</td>
<td>70%</td>
<td>paf_threshold = 20 pb_threshold = .75 window = 10 tau = 6</td>
</tr>
</tbody>
</table>

4.2 Approximate Entropy and Discrete-output HMM

Using the technique developed by Yuan Qi, the inter-beat interval time series in the CinC data sets were converted into an entropy series, and discretized for HMM analysis.
The level of discretization was chosen to be 100, which resulted in as little information loss as possible (Fig 10).

Figure 10 - discretized entropy series, with enough discrete levels to minimize information loss

![Entrophy Series](image)

The entropy series were then analyzed using discrete-output HMMs with the results in the following table. The results were less than satisfactory given the low predictive accuracy, and histogram plots (Fig 6) of the entropy data were examined for visible features that would allow for discrimination between the population groups. No extraordinary differing factors could be found. Another remedy that was attempted was adjusting the resolution of the spectrum estimation, also with no obvious improvement.

<table>
<thead>
<tr>
<th>CinCSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Pred Acc</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>21</td>
<td>3</td>
<td>19</td>
<td>7</td>
<td>75%</td>
<td>14%</td>
<td>48%</td>
<td>Q = 2</td>
</tr>
<tr>
<td>P vs. N</td>
<td>21</td>
<td>2</td>
<td>20</td>
<td>7</td>
<td>75%</td>
<td>9%</td>
<td>46%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>P vs. N</td>
<td>21</td>
<td>3</td>
<td>19</td>
<td>7</td>
<td>75%</td>
<td>14%</td>
<td>48%</td>
<td>Q = 4</td>
</tr>
<tr>
<td>P vs. N</td>
<td>18</td>
<td>5</td>
<td>17</td>
<td>10</td>
<td>64%</td>
<td>23%</td>
<td>46%</td>
<td>Q = 15</td>
</tr>
</tbody>
</table>

Table 5 - O (number of discrete levels used) was 100, trained on Groups P & N and tested on CinC dataset for Event 1 Screening

4.3 Gaussian Mixture HMM on IBIs

When the results using the entropy did not meet expectations, the decision was made to focus instead on IBIs. Histogram plots of IBIs over different time segments revealed
some interesting patterns (Fig 6), notably the tendency of IBIs to diverge into clusters of long IBIs and short IBIs. This observation supports the initial observation that PBs become more prevalent before and during PAF. Even though histograms were useful in identifying this pattern of IBI change over time, it doesn’t capture the pattern of IBI changes over time. As a result, the IBI series were analyzed using Gaussian-mixture HMMs, which would capture the both the ectopic events, but also additional information on timing and frequency.

The results for the analysis are presented below in the order of Events attempted.

4.3.1 Event 1: Screening

Table 6

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>88%</td>
<td>54%</td>
<td>72%</td>
<td>M =2</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>14</td>
<td>22</td>
<td>2</td>
<td>2</td>
<td>88%</td>
<td>92%</td>
<td>90%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>4</td>
<td>12</td>
<td>12</td>
<td>9</td>
<td>31%</td>
<td>50%</td>
<td>43%</td>
<td></td>
</tr>
</tbody>
</table>

*best predictive accuracy for both P vs. N and P vs. Q

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>14</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>86%</td>
<td>38%</td>
<td>66%</td>
<td>M =3</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>16</td>
<td>18</td>
<td>6</td>
<td>0</td>
<td>100%</td>
<td>75%</td>
<td>85%</td>
<td>Q = 2</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>6</td>
<td>6</td>
<td>18</td>
<td>7</td>
<td>46%</td>
<td>25%</td>
<td>32%</td>
<td></td>
</tr>
</tbody>
</table>

*highest P vs. Q sensitivity

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>16</td>
<td>11</td>
<td>9</td>
<td>2</td>
<td>89%</td>
<td>55%</td>
<td>71%</td>
<td>M =2</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>16</td>
<td>22</td>
<td>5</td>
<td>2</td>
<td>89%</td>
<td>81%</td>
<td>84%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>11</td>
<td>45%</td>
<td>44%</td>
<td>45%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CinCSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>27</td>
<td>8</td>
<td>14</td>
<td>1</td>
<td>96%</td>
<td>57%</td>
<td>70%</td>
<td>M =2, Q = 1</td>
</tr>
</tbody>
</table>

*parameters of best predictive accuracy on the evaluation set were used on the test set and the CinC set
### 4.3.2 Event 2: Detection

#### Table 7

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>11</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>100%</td>
<td>79%</td>
<td>88%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>8</td>
<td>33</td>
<td>2</td>
<td>3</td>
<td>73</td>
<td>94</td>
<td>89</td>
<td>Q = 3</td>
</tr>
<tr>
<td>Ne vs. Qc</td>
<td>7</td>
<td>29</td>
<td>6</td>
<td>7</td>
<td>50</td>
<td>83</td>
<td>73</td>
<td></td>
</tr>
</tbody>
</table>

*highest predictive accuracy for P2c vs. Nc

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>7</td>
<td>14</td>
<td>0</td>
<td>4</td>
<td>64%</td>
<td>100%</td>
<td>84%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>8</td>
<td>33</td>
<td>2</td>
<td>3</td>
<td>73</td>
<td>94</td>
<td>89</td>
<td>Q = 2</td>
</tr>
<tr>
<td>Ne vs. Qc</td>
<td>9</td>
<td>26</td>
<td>9</td>
<td>5</td>
<td>64</td>
<td>74</td>
<td>71</td>
<td></td>
</tr>
</tbody>
</table>

*highest specificity for P2c vs. Nc

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs Nc</td>
<td>11</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>100%</td>
<td>79%</td>
<td>88%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>11</td>
<td>32</td>
<td>3</td>
<td>0</td>
<td>100%</td>
<td>91</td>
<td>93</td>
<td>Q = 1</td>
</tr>
<tr>
<td>Ne vs. Qc</td>
<td>9</td>
<td>23</td>
<td>12</td>
<td>5</td>
<td>64</td>
<td>65</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>

*highest predictive accuracy for P2c vs. Qc

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>11</td>
<td>11</td>
<td>6</td>
<td>0</td>
<td>100%</td>
<td>65%</td>
<td>79%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>9</td>
<td>21</td>
<td>6</td>
<td>0</td>
<td>100%</td>
<td>78%</td>
<td>83%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>Ne vs. Qc</td>
<td>1</td>
<td>25</td>
<td>0</td>
<td>16</td>
<td>6%</td>
<td>100%</td>
<td>62%</td>
<td></td>
</tr>
</tbody>
</table>

*parameters of best predictive accuracy for P2c vs. Nc used for test set

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. P1c</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>88%</td>
<td>64%</td>
<td>74%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>8</td>
<td>21</td>
<td>6</td>
<td>0</td>
<td>100%</td>
<td>78%</td>
<td>83%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>P1c vs. Qc</td>
<td>9</td>
<td>21</td>
<td>6</td>
<td>2</td>
<td>82%</td>
<td>78%</td>
<td>79%</td>
<td></td>
</tr>
</tbody>
</table>

*best predictive accuracy for P2c vs. P1c

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. P1c</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>89%</td>
<td>71%</td>
<td>81%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>9</td>
<td>21</td>
<td>6</td>
<td>0</td>
<td>100%</td>
<td>78%</td>
<td>83%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>P1c vs. Qc</td>
<td>7</td>
<td>19</td>
<td>8</td>
<td>0</td>
<td>100%</td>
<td>70%</td>
<td>76%</td>
<td></td>
</tr>
</tbody>
</table>

*parameters of best predictive accuracy for P2c vs. P1c used for test set
4.3.3 Event 3: Prediction

Table 8

Using six 200 IBI segments:

<table>
<thead>
<tr>
<th></th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 vs. P2</td>
<td>11</td>
<td>22</td>
<td>3</td>
<td>14</td>
<td>44%</td>
<td>88%</td>
<td>66%</td>
<td>M = 1</td>
</tr>
<tr>
<td>CinC</td>
<td>7</td>
<td>61</td>
<td>11</td>
<td>21</td>
<td>25%</td>
<td>85%</td>
<td>68%</td>
<td>Q = 1</td>
</tr>
</tbody>
</table>

Using 1200 IBI segments (similar to Event 1):

<table>
<thead>
<tr>
<th></th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>EvalSet P2 vs. P1</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>50%</td>
<td>75%</td>
<td>63%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2 vs. Q</td>
<td>5</td>
<td>27</td>
<td>3</td>
<td>3</td>
<td>63%</td>
<td>90%</td>
<td>84%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>P1 vs. Q</td>
<td>7</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td>88%</td>
<td>83%</td>
<td>84%</td>
<td></td>
</tr>
</tbody>
</table>

*highest predictive accuracy for P2 vs. P1 and P2 vs. Q

<table>
<thead>
<tr>
<th></th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>TestSet P2 vs. P1</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>45%</td>
<td>75%</td>
<td>58%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2 vs. Q</td>
<td>9</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td>82%</td>
<td>87%</td>
<td>85%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>P1 vs. Q</td>
<td>6</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td>75%</td>
<td>87%</td>
<td>84%</td>
<td></td>
</tr>
</tbody>
</table>

*parameters used on TestSet and CinCSet

4.4 Gaussian Mixture IMMs on Spectrogram

Event 1: Screening
(Note: number of frequency bins set to 50)

Table 9

<table>
<thead>
<tr>
<th></th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>EvalSet P vs. N</td>
<td>2</td>
<td>28</td>
<td>0</td>
<td>16</td>
<td>11%</td>
<td>100%</td>
<td>65%</td>
<td>M = 5, Q = 1</td>
</tr>
</tbody>
</table>

*best predictive accuracy and sensitivity

<table>
<thead>
<tr>
<th></th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>TestSet P vs. N</td>
<td>2</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>12%</td>
<td>100%</td>
<td>53%</td>
<td>M = 5, Q = 1</td>
</tr>
</tbody>
</table>
Event 1: Screening  
(Note: number of frequency bins set to 20)  
Table 10

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>13</td>
<td>4</td>
<td>13</td>
<td>2</td>
<td>87%</td>
<td>24%</td>
<td>53%</td>
<td>M = 5, Q = 1</td>
</tr>
</tbody>
</table>

*best predictive accuracy and sensitivity

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. Q</td>
<td>8</td>
<td>26</td>
<td>2</td>
<td>6</td>
<td>57%</td>
<td>93%</td>
<td>81%</td>
<td>M = 1, Q = 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>13</td>
<td>4</td>
<td>11</td>
<td>1</td>
<td>93%</td>
<td>27%</td>
<td>59%</td>
<td>M = 5, Q = 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. Q</td>
<td>7</td>
<td>21</td>
<td>1</td>
<td>8</td>
<td>53%</td>
<td>95%</td>
<td>78%</td>
<td>M = 1, Q = 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CinCSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>10</td>
<td>4</td>
<td>8</td>
<td>3</td>
<td>77%</td>
<td>33%</td>
<td>56%</td>
<td>M = 5, Q = 1</td>
</tr>
</tbody>
</table>

*best parameters used on testset and CinCSet

Event 2: Detection  
(Note: number of frequency bins set to 20)  
Table 11

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>1</td>
<td>17</td>
<td>2</td>
<td>7</td>
<td>13%</td>
<td>89%</td>
<td>67%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>2</td>
<td>30</td>
<td>0</td>
<td>11</td>
<td>15%</td>
<td>100%</td>
<td>74%</td>
<td>Q = 2</td>
</tr>
</tbody>
</table>

*best predictive accuracy and sensitivity

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>10</td>
<td>9%</td>
<td>94%</td>
<td>61%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>15</td>
<td>6%</td>
<td>100%</td>
<td>44%</td>
<td>Q = 2</td>
</tr>
</tbody>
</table>

*best parameters used on testset

Event 3: Prediction  
(Note: number of frequency bins set to 20)  
Table 12

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2 vs. N</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>27%</td>
<td>67%</td>
<td>50%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P2 vs. Q</td>
<td>2</td>
<td>24</td>
<td>7</td>
<td>4</td>
<td>33%</td>
<td>77%</td>
<td>70%</td>
<td>Q = 1</td>
</tr>
</tbody>
</table>

*best predictive accuracy and sensitivity
<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2 vs. N</td>
<td>3</td>
<td>15</td>
<td>5</td>
<td>4</td>
<td>43%</td>
<td>75%</td>
<td>67%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P2 vs. Q</td>
<td>3</td>
<td>15</td>
<td>3</td>
<td>7</td>
<td>30%</td>
<td>83%</td>
<td>64%</td>
<td>Q = 1</td>
</tr>
</tbody>
</table>

*best parameters used on testset

5.0 Discussion

5.1 Event 1: Screening

In the evaluation sets, the technique of using Gaussian mixture modeling in Event 1 demonstrated significant specificity (True Neg/ [True Neg + False Pos]) ranging up to 92% and high sensitivity (True Pos/ [True Pos + False Neg]) ranging up to 100% in comparison between Group P and Group Q. The results were comparatively lower between those of Group P and Group N, with sensitivity high of 94% and specificity high of 54%. This makes some sense since one would expect that there is greater difference between patients with PAF and those who are healthy, as opposed to comparison between those with PAF and those who might have some unspecified arrhythmia. A comparison between Group N and Group Q resulted in only 46% sensitivity (screening for N-types) and 92% specificity. This could be due to the fact that some of those in Group N have no arrhythmic disorders. The test set for Event 1 verified the general distribution of results shown in the evaluation sets, using one state of 2 Gaussians as a model for the system. The final test was for discrimination between the set of Group Ns and Ps provided by the original CinC test set. The Gaussian-mixture technique provided sensitivity of 96% and a specificity of 57%, with an overall accuracy of 70%. What would make this even more promising is if the Gaussian-mixture outputs were used in conjunction with the Prematurity Weighting method, which has a higher specificity than it does sensitivity.
Testing for Event 1 using discrete-output HMMs on entropy yielded disappointing results by comparison. The sensitivity had a high of 75%, but had only a specificity of 23%. Using Gaussian-mixture HMMs on spectrograms with a frequency resolution of 50 bins yielded an accuracy high of 12% in the evaluation sets, and a specificity around 100%. The additional testing for Event 1 using Gaussian-mixture HMMs on spectrograms with a frequency resolution of 20 bins yielded much better results, with an accuracy high of 57% in the evaluation sets, and a specificity around 93%. Applying the algorithm to the CinC test set resulted in only a 56% prediction accuracy.

5.2 Event 2 : Detection

In testing for Event 2, the algorithm for Gaussian-mixture HMMs on IBIs was used. The results from the evaluation sets were very promising, with a high of 100% accuracy in recognizing the immediate onset of PAF and up to 100% specificity depending on the parameters used. Since actual periods of PAF were not included in the CinC test set, the only one-time test was given based on the randomized test set. The result was an 89% accuracy of discrimination between PAF and non-PAF episodes. Given that PAF is an atrial disorder, and the technique used for detection is only utilizing the time between QRS complexes (a primarily ventricular event), these results further demonstrates the importance of IBIs as an overall indicator of cardiac health, as opposed to localized ventricular state.
5.3 Event 3: Prediction

For Event 3, the results from the segmented HMM analysis (200 IBI blocks) were the reverse of what was observed in Event 2, with a high specificity and a relatively low sensitivity. Interestingly, when tested against the test set provided in the original CinC competition, the technique returned a result of 68% predictive accuracy, but with a much higher specificity (86%) than sensitivity (25%), which would not have translated into a high score in the competition. In hindsight, the test could have been better structured given more training data for the healthy state for an individual, as the training data might have been insufficient. Lack of data preclude further exploration in this study.

By comparison, the use of HMMs directly trained on the entire 1200 IBIs yielded better sensitivity (64% on the CinC set) and a better scoring on the CinC testing (18/22). Also noteworthy is the fact that when P-2 patients are trained and tested against type Q patients, the sensitivity and specificity both jump to be around 90%.

5.4 Entropy Failure Speculation

The failure of the entropy analysis was a major surprise given previous literature and experiments [Vikman et al, 1999], which suggested that approximate entropy (ApEn) would decrease prior to cardiac episodes. ApEn measures the logarithmic likelihood that runs of patterns that are close to each other will remain close in the subsequent incremental comparisons. A time series containing many repetitive patterns has a relatively small ApEn; conversely, more random data produce higher values. The cause of failure with entropy is uncertain, but possibilities include either mismatched resolutions, or possible discrepancy in each patient due to emotional or physical distress
not documented at the time of recording. Whereas the variability in the power spectrum might not be sufficient to characterize PAF onset, the spectrogram itself offered two features worth considering. The first was the vertical streaking corresponding to ectopic beats, and the other the shifting of control frequency (relative strengths of sympathetic, parasympathetic output) due to ectopic behavior. The fact that direct training on the spectrogram yielded better results is encouraging, though further analysis utilizing tradeoffs between time resolution and frequency resolution could potentially prove even more enlightening.

6.0 Conclusions

This study demonstrates the feasibility of screening, detection, and prediction of PAF based solely on the pattern of inter-beat intervals. Features extracted from IBIs, such as entropy and spectrograms were also explored as possible discriminators. The application of pattern recognition techniques builds on the previous work with threshold and weighting functions for narrowly defined events. Feasibility of using Gaussian-mixture HMMs in PAF analysis yielded unambiguous results. Especially noteworthy was the results regarding comparisons between those with PAF and healthy patients. While the techniques appear robust enough to differentiate between PAF and non-PAF patients with only reasonable accuracy, it is highly successful (~90% sensitivity and specificity) when used in discriminating between healthy individuals and those with PAF.
7.0 Recommended Future Works

The key application of IBI-based PAF prediction lies in surface-lead ECG processing, where the noise from muscular distortions and lead contact movements contribute to a fundamental limit in the accurate tracking of small amplitude P-waves. By focusing on the larger and more distinctive QRS complexes, the techniques analyzed in this study are better suited to the type of commercial applications in the area of health and fitness products. There are several specific areas in the determination of health state that this author would like to see explored in the future.

In terms of continuing research in use of IBIs in PAF analysis, the patient-specific training that was described in Section 3.3.3 bears great potential. The ultimate goal would be to enable devices to adapt to personal variability in physiology and provide prediction measures based on individual case history. To achieve this type of analysis, what would most likely be required is longer ECGs of PAF patients than what was used in this study. Training could then be applied to regions distal and in proximity to PAF episodes for that individual.

In terms of academic questions left to explore, there remains the question of screening and prediction for other atrial arrhythmias. Are IBIs equally important in identifying other arrhythmic episodes? Are these techniques valid in the presence of artificial stimuli, such as drugs? Once prediction accuracies could be established to within a certain threshold, there would also be many questions that would follow regarding measurement of the exact preventative value for interventions, in terms of successfully prevented arrhythmic episodes.
Besides academic pursuits that could draw on this research, there are many potential uses of this technology to benefit public health and services. The prevalence of PAF in the general population has never been accurately determined, and this study could contribute much to making available to the public the necessary testing and evaluation that would otherwise require access to a medical institution. Rather than relying on skilled technicians, expensive hardware, and human expert interpretation, the algorithms could be easily automated and calibrated to work with commercially available fitness sensors. The ideas are currently being explored in various institutions. Among commercial entities, Motorola in particular has taken a strong interest in incorporating algorithms into mobile devices for providing valuable service to their customers. Continuation along the current paths of research would be likely to yield great future dividends in terms of quality health care that is both personal and pervasive.
Appendix A

Complete HMM Results

Gaussian-Mixture HMMs on IBIs

Event 1: Screening

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>15</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>94%</td>
<td>39%</td>
<td>69%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>14</td>
<td>19</td>
<td>5</td>
<td>2</td>
<td>86%</td>
<td>79%</td>
<td>83%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>4</td>
<td>9</td>
<td>15</td>
<td>9</td>
<td>31%</td>
<td>38%</td>
<td>35%</td>
<td></td>
</tr>
</tbody>
</table>

Toolkit commandlines used to generate above:

- [priorli, transmatli, mixmatli, muli, Sigmal] = init_mhmm(mod_tri, 1, 3, 'diag', 0)
- [LL, priorli, transmatli, muli, Sigmal, mixmatli] = learn_mhmm(mod_tri, priorli, transmatli, muli, Sigmal, mixmatli, 5);
- loglikl = log_lik_mhmm(mod_tri, priorl, transmatli, mixmatli, mul, Sigmal)

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>88%</td>
<td>54%</td>
<td>72%</td>
<td>M = 2</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>14</td>
<td>22</td>
<td>2</td>
<td>2</td>
<td>88%</td>
<td>92%</td>
<td>90%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>4</td>
<td>12</td>
<td>12</td>
<td>9</td>
<td>31%</td>
<td>50%</td>
<td>43%</td>
<td></td>
</tr>
</tbody>
</table>

Toolkit commandlines used to generate above:

- [priorli, transmatli, mixmatli, muli, Sigmal] = init_mhmm(mod_tri, 1, 2, 'diag', 0)
- [LL, priorli, transmatli, muli, Sigmal, mixmatli] = learn_mhmm(mod_tri, priorli, transmatli, muli, Sigmal, mixmatli, 5);
- loglikl = log_lik_mhmm(mod_tri, priorl, transmatli, mixmatli, mul, Sigmal)

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>10</td>
<td>3</td>
<td>10</td>
<td>6</td>
<td>63%</td>
<td>23%</td>
<td>45%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>50%</td>
<td>67%</td>
<td>60%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>3</td>
<td>22</td>
<td>2</td>
<td>10</td>
<td>23%</td>
<td>92%</td>
<td>68%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>15</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>94%</td>
<td>38%</td>
<td>69%</td>
<td>M = 5</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>15</td>
<td>18</td>
<td>6</td>
<td>1</td>
<td>94%</td>
<td>75%</td>
<td>83%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>4</td>
<td>12</td>
<td>12</td>
<td>9</td>
<td>31%</td>
<td>50%</td>
<td>43%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>14</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>86%</td>
<td>38%</td>
<td>66%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>16</td>
<td>18</td>
<td>6</td>
<td>0</td>
<td>100%</td>
<td>75%</td>
<td>85%</td>
<td>Q = 2</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>6</td>
<td>6</td>
<td>18</td>
<td>7</td>
<td>46%</td>
<td>25%</td>
<td>32%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>15</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>94%</td>
<td>23%</td>
<td>62%</td>
<td>M = 2</td>
</tr>
<tr>
<td>P vs. Q</td>
<td>16</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100%</td>
<td>50%</td>
<td>70%</td>
<td>Q = 2</td>
</tr>
<tr>
<td>N vs. Q</td>
<td>5</td>
<td>11</td>
<td>13</td>
<td>8</td>
<td>38%</td>
<td>46%</td>
<td>43%</td>
<td>left-right</td>
</tr>
<tr>
<td>EvalSet</td>
<td>TruePos</td>
<td>TrueNeg</td>
<td>FalsePos</td>
<td>FalseNeg</td>
<td>Sens</td>
<td>Spec</td>
<td>Pred</td>
<td>Parameters</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>P vs. N</strong></td>
<td>14</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>86%</td>
<td>38%</td>
<td>66%</td>
<td>M = 2</td>
</tr>
<tr>
<td><strong>P vs. Q</strong></td>
<td>13</td>
<td>19</td>
<td>5</td>
<td>3</td>
<td>81%</td>
<td>79%</td>
<td>80%</td>
<td>Q = 2</td>
</tr>
<tr>
<td><strong>N vs. Q</strong></td>
<td>4</td>
<td>14</td>
<td>10</td>
<td>9</td>
<td>31%</td>
<td>58%</td>
<td>49%</td>
<td>left-right</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>EvalSet</strong></th>
<th><strong>TruePos</strong></th>
<th><strong>TrueNeg</strong></th>
<th><strong>FalsePos</strong></th>
<th><strong>FalseNeg</strong></th>
<th><strong>Sens</strong></th>
<th><strong>Spec</strong></th>
<th><strong>Pred</strong></th>
<th><strong>Parameters</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P vs. N</strong></td>
<td>13</td>
<td>1</td>
<td>12</td>
<td>3</td>
<td>81%</td>
<td>8%</td>
<td>48%</td>
<td>M = 1</td>
</tr>
<tr>
<td><strong>P vs. Q</strong></td>
<td>15</td>
<td>2</td>
<td>22</td>
<td>1</td>
<td>94%</td>
<td>8%</td>
<td>43%</td>
<td>Q = 2</td>
</tr>
<tr>
<td><strong>N vs. Q</strong></td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>9</td>
<td>31%</td>
<td>33%</td>
<td>32%</td>
<td>left-right</td>
</tr>
</tbody>
</table>

Toolkit commandlines used to generate above:

\[
\text{[priorli, transmatli, mixmatli, muli, Sigmali]} = \text{init_mhmm(mod_tri, 2, 1, 'diag', 1)}
\]

\[
\text{[LL, priorli, transmatli, muli, Sigmali]} = \text{learn_mhmm(mod_tri, priorli, transmatli, muli, Sigmali, mixmatli, 5)};
\]

\[
\text{loglikl} = \text{log_lik mhmm(mod_tri, prior, transmat, mixmat, mul, Sigma)}
\]

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P vs. N</strong></td>
<td>16</td>
<td>11</td>
<td>9</td>
<td>2</td>
<td>89%</td>
<td>55%</td>
<td>71%</td>
<td>M = 2</td>
</tr>
<tr>
<td><strong>P vs. Q</strong></td>
<td>16</td>
<td>22</td>
<td>5</td>
<td>2</td>
<td>89%</td>
<td>81%</td>
<td>84%</td>
<td>Q = 1</td>
</tr>
<tr>
<td><strong>N vs. Q</strong></td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>11</td>
<td>45%</td>
<td>44%</td>
<td>45%</td>
<td>left-right</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P2c vs. Nc</strong></td>
<td>10</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>91%</td>
<td>93%</td>
<td>92%</td>
<td>M = 1</td>
</tr>
<tr>
<td><strong>P2c vs. Qc</strong></td>
<td>8</td>
<td>33</td>
<td>2</td>
<td>3</td>
<td>73%</td>
<td>94%</td>
<td>89%</td>
<td>Q = 3</td>
</tr>
<tr>
<td><strong>Nc vs. Qc</strong></td>
<td>7</td>
<td>29</td>
<td>6</td>
<td>7</td>
<td>50%</td>
<td>83%</td>
<td>73%</td>
<td>left-right</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P2c vs Nc</strong></td>
<td>11</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>100%</td>
<td>79%</td>
<td>88%</td>
<td>M = 1</td>
</tr>
<tr>
<td><strong>P2c vs. Qc</strong></td>
<td>11</td>
<td>32</td>
<td>3</td>
<td>0</td>
<td>100%</td>
<td>91%</td>
<td>93%</td>
<td>Q = 3</td>
</tr>
<tr>
<td><strong>Nc vs. Qc</strong></td>
<td>3</td>
<td>32</td>
<td>3</td>
<td>11</td>
<td>21%</td>
<td>91%</td>
<td>71%</td>
<td>left-right</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P2c vs Nc</strong></td>
<td>8</td>
<td>14</td>
<td>0</td>
<td>3</td>
<td>73%</td>
<td>100%</td>
<td>88%</td>
<td>M = 1</td>
</tr>
<tr>
<td><strong>P2c vs. Qc</strong></td>
<td>8</td>
<td>33</td>
<td>2</td>
<td>3</td>
<td>73%</td>
<td>94%</td>
<td>89%</td>
<td>Q = 2</td>
</tr>
<tr>
<td><strong>Nc vs. Qc</strong></td>
<td>9</td>
<td>27</td>
<td>8</td>
<td>5</td>
<td>64%</td>
<td>77%</td>
<td>73%</td>
<td>left-right</td>
</tr>
<tr>
<td>EvalSet</td>
<td>TruePos</td>
<td>TrueNeg</td>
<td>FalsePos</td>
<td>FalseNeg</td>
<td>Sens</td>
<td>Spec</td>
<td>Pred</td>
<td>Parameters</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>P2c vs Nc</td>
<td>7</td>
<td>14</td>
<td>0</td>
<td>4</td>
<td>64%</td>
<td>100%</td>
<td>84%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>8</td>
<td>33</td>
<td>2</td>
<td>3</td>
<td>73%</td>
<td>94%</td>
<td>89%</td>
<td>Q = 2</td>
</tr>
<tr>
<td>Nc vs. Qc</td>
<td>9</td>
<td>26</td>
<td>9</td>
<td>5</td>
<td>64%</td>
<td>74%</td>
<td>71%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs Nc</td>
<td>11</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>100%</td>
<td>79%</td>
<td>88%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>11</td>
<td>32</td>
<td>3</td>
<td>0</td>
<td>100%</td>
<td>91%</td>
<td>94%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>Nc vs. Qc</td>
<td>9</td>
<td>23</td>
<td>12</td>
<td>5</td>
<td>64%</td>
<td>65%</td>
<td>64%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs Nc</td>
<td>7</td>
<td>13</td>
<td>1</td>
<td>4</td>
<td>64%</td>
<td>93%</td>
<td>80%</td>
<td>M = 2</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>7</td>
<td>35</td>
<td>0</td>
<td>4</td>
<td>64%</td>
<td>100%</td>
<td>91%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>Nc vs. Qc</td>
<td>9</td>
<td>27</td>
<td>8</td>
<td>5</td>
<td>64%</td>
<td>77%</td>
<td>73%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs Nc</td>
<td>8</td>
<td>13</td>
<td>1</td>
<td>3</td>
<td>73%</td>
<td>93%</td>
<td>84%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>7</td>
<td>34</td>
<td>1</td>
<td>4</td>
<td>64%</td>
<td>97%</td>
<td>89%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>Nc vs. Qc</td>
<td>11</td>
<td>26</td>
<td>9</td>
<td>3</td>
<td>79%</td>
<td>74%</td>
<td>76%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>1</td>
<td>16</td>
<td>0</td>
<td>2</td>
<td>33%</td>
<td>100%</td>
<td>89%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>1</td>
<td>23</td>
<td>1</td>
<td>2</td>
<td>33%</td>
<td>96%</td>
<td>89%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>Nc vs. Qc</td>
<td>7</td>
<td>22</td>
<td>2</td>
<td>9</td>
<td>44%</td>
<td>92%</td>
<td>73%</td>
<td></td>
</tr>
</tbody>
</table>

Sequence Re-run to generate a larger test set of P2c’s:

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs Nc</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td>0</td>
<td>100%</td>
<td>63%</td>
<td>74%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>8</td>
<td>23</td>
<td>2</td>
<td>3</td>
<td>73%</td>
<td>92%</td>
<td>86%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>Nc vs. Qc</td>
<td>2</td>
<td>26</td>
<td>0</td>
<td>17</td>
<td>11%</td>
<td>100%</td>
<td>63%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>11</td>
<td>11</td>
<td>6</td>
<td>0</td>
<td>100%</td>
<td>65%</td>
<td>79%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>11</td>
<td>21</td>
<td>4</td>
<td>2</td>
<td>85%</td>
<td>84%</td>
<td>84%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>Nc vs. Qc</td>
<td>1</td>
<td>25</td>
<td>0</td>
<td>16</td>
<td>6%</td>
<td>100%</td>
<td>62%</td>
<td></td>
</tr>
</tbody>
</table>
### Event 3: Prediction

Using six 200 IBI segments:

<table>
<thead>
<tr>
<th></th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 vs. P2</td>
<td>11</td>
<td>22</td>
<td>3</td>
<td>14</td>
<td>44%</td>
<td>88%</td>
<td>66%</td>
<td>M = 1</td>
</tr>
<tr>
<td>CinC</td>
<td>7</td>
<td>61</td>
<td>11</td>
<td>21</td>
<td>25%</td>
<td>85%</td>
<td>68%</td>
<td>Q = 1</td>
</tr>
</tbody>
</table>

Using 1200 IBI segments (similar to Event 1):

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2-P1</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>50%</td>
<td>75%</td>
<td>63%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P2-Q</td>
<td>5</td>
<td>27</td>
<td>3</td>
<td>3</td>
<td>63%</td>
<td>90%</td>
<td>84%</td>
<td>Q = 1</td>
</tr>
<tr>
<td>P1-Q</td>
<td>6</td>
<td>28</td>
<td>2</td>
<td>2</td>
<td>75%</td>
<td>93%</td>
<td>89%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2-P1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>38%</td>
<td>88%</td>
<td>63%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2-Q</td>
<td>6</td>
<td>27</td>
<td>3</td>
<td>2</td>
<td>75%</td>
<td>90%</td>
<td>87%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>P1-Q</td>
<td>7</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td>88%</td>
<td>83%</td>
<td>84%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2-P1</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>55%</td>
<td>75%</td>
<td>63%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2-Q</td>
<td>9</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td>82%</td>
<td>87%</td>
<td>85%</td>
<td>Q = 3</td>
</tr>
<tr>
<td>P1-Q</td>
<td>6</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td>75%</td>
<td>87%</td>
<td>84%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CinCSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2 vs. P1</td>
<td>18</td>
<td>41</td>
<td>31</td>
<td>10</td>
<td>64%</td>
<td>57%</td>
<td>59%</td>
<td>M = 1, Q = 3</td>
</tr>
</tbody>
</table>

### Gaussian Mixture HMMs on Spectrogram

#### Event 1: Screening

*Note: number of frequency bins set to 20*

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>13</td>
<td>4</td>
<td>13</td>
<td>2</td>
<td>87%</td>
<td>24%</td>
<td>53%</td>
<td>M = 5, Q = 1</td>
</tr>
</tbody>
</table>

Toolkit commandlines used to generate above:

- \[
  \text{[priorli, transmatli, mixmatli, muli, Sigmal]} = \text{initmhmm(modtri, 1, 5, 'full', 0)}
  \]
- \[
  \text{[LL, priorli, transmatli, muli, Sigmal, mixmatli]} = \text{learnmhmm(modtri, priorli, transmatli, muli, Sigmal, mixmatli, 5)};
  \]
- \[
  \text{loglikl} = \text{loglikmhmm(modtri, priorli, transmatli, mixmatli, muli, Sigmal)}
  \]

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. N</td>
<td>12</td>
<td>4</td>
<td>11</td>
<td>2</td>
<td>86%</td>
<td>27%</td>
<td>55%</td>
<td>M = 5, Q = 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TestSet</td>
<td>TruePos</td>
<td>TrueNeg</td>
<td>FalsePos</td>
<td>FalseNeg</td>
<td>Sens</td>
<td>Spec</td>
<td>Pred</td>
<td>Parameters</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>P vs. N</td>
<td>13</td>
<td>4</td>
<td>11</td>
<td>1</td>
<td>93%</td>
<td>27%</td>
<td>59%</td>
<td>M = 5, Q = 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. Q</td>
<td>7</td>
<td>21</td>
<td>1</td>
<td>8</td>
<td>53%</td>
<td>95%</td>
<td>78%</td>
<td>M = 1, Q = 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P vs. Q</td>
<td>7</td>
<td>21</td>
<td>1</td>
<td>8</td>
<td>53%</td>
<td>95%</td>
<td>78%</td>
<td>M = 1, Q = 2</td>
</tr>
</tbody>
</table>

**Event 2: Detection**
*Note: number of frequency bins set to 20*

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>1</td>
<td>17</td>
<td>2</td>
<td>7</td>
<td>13%</td>
<td>89%</td>
<td>67%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>2</td>
<td>30</td>
<td>0</td>
<td>11</td>
<td>15%</td>
<td>100%</td>
<td>74%</td>
<td>Q = 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2c vs. Nc</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>10</td>
<td>9%</td>
<td>94%</td>
<td>61%</td>
<td>M = 1</td>
</tr>
<tr>
<td>P2c vs. Qc</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>15</td>
<td>6%</td>
<td>100%</td>
<td>44%</td>
<td>Q = 2</td>
</tr>
</tbody>
</table>

**Event 3: Prediction**
*Note: number of frequency bins set to 20*

<table>
<thead>
<tr>
<th>EvalSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2 vs. N</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>27%</td>
<td>67%</td>
<td>50%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P2 vs. Q</td>
<td>2</td>
<td>24</td>
<td>7</td>
<td>4</td>
<td>33%</td>
<td>77%</td>
<td>70%</td>
<td>Q = 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestSet</th>
<th>TruePos</th>
<th>TrueNeg</th>
<th>FalsePos</th>
<th>FalseNeg</th>
<th>Sens</th>
<th>Spec</th>
<th>Pred</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2 vs. N</td>
<td>3</td>
<td>15</td>
<td>5</td>
<td>4</td>
<td>43%</td>
<td>75%</td>
<td>67%</td>
<td>M = 3</td>
</tr>
<tr>
<td>P2 vs. Q</td>
<td>3</td>
<td>15</td>
<td>3</td>
<td>7</td>
<td>30%</td>
<td>83%</td>
<td>64%</td>
<td>Q = 1</td>
</tr>
</tbody>
</table>
Appendix B

Sample Spectrograms (frequency resolution of 20 bins)

Spectrograms of Type P (8 randomly selected):

(a)  
(b)  
(c)  
(d)  

The four spectrograms above demonstrate expected behavior: there is frequent ectopic activity observable. The two spectrograms on the right are preceding PAF episodes (type P-2), and appear to show greater frequency of ectopic activity towards the end. The two on the left are not preceding PAF episodes (type P-1), and have scattered ectopic activity. All spectrograms are followed by the spectrogram of its continuation set (either showing PAF, or not), as separated by the dark blue line.

These two spectrograms (left, P-1; right, P-2) appear strangely quiescent for a patient diagnosed with PAF.
Relatively quiescent, with events near the end, suggesting precursor to PAF episode, but is actually distant from any episode, type P-1.

Type P-2 with strong ectopic behavior long before a reported episode. Visual inspection of ECG revealed premature beats, bursts of tachycardia, and indeed no PAF prior to the continuation set.

Spectrograms of Type N (8 randomly selected):

These five spectrograms of non-PAF patients appears relatively healthy. There are few visible ectopic beats.
These non-PAF patients have spectrograms that demonstrate ectopic behavior.

Spectrograms of Type Q (8 randomly selected):

(a)  
(b)  
(c)  
(d)
All eight spectrums of type Q exhibit only minor irregularities, if any.
Appendix C

Sample Spectrograms (frequency resolution of 50 bins)

Spectrograms of Type P (same order as presented in Appendix B):

(a)

(b)

(c)

(d)

Note the smearing effect in time compared to spectrograms with 20 frequency bins (Appendix B). On the bottom two spectrograms, greater frequency resolution could be observed in the horizontal bands.

(e)

(f)
Type P-1.

Type P-2 with strong ectopic behavior long before a reported episode, serious smearing in time.

Spectrograms of Type N (8 randomly selected):

These five spectrograms of non-PAF patients appears relatively healthy. There are few visible ectopic beats.
These non-PAF patients have spectrograms that demonstrate ectopic behavior.
References


