Chagas Prediction Using ECG Data with Dynamic Bayesian Networks

Chagas Disease

Chagas is a tropical parasitic illness.

Acute Phase: Often mild or asymptomatic, lasts ~8–12 weeks.

Chronic Phase: Can remain silent for years, but ~30% develop chronic Chagas cardiomyopathy (CCC).

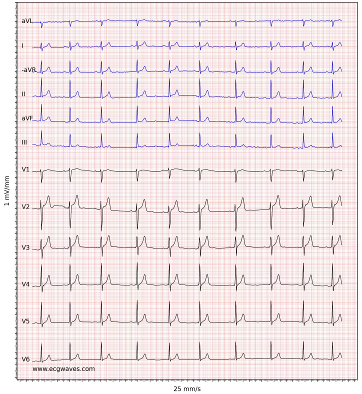
What is ECG Data and why use it?

Records the heart's electrical activity from **12 different perspectives**, using **10 electrodes** placed on the chest and limbs.

Chagas affects the heart’s structure and rhythm:

These abnormalities are **detectable via ECG**, making ECG-based prediction crucial for early diagnosis and treatment planning.

Chagas-induced changes in **conduction patterns**, **wave durations**, and **rhythm irregularities** are captured across different leads and time intervals.



Dataset

PhysioNet Challenge on Chagas Disease Detection

- Provides 3 different datasets of ECG Data:
  - One small (1,600 samples) all positive cases (strongly labeled)
  - Two larger datasets (300,000 and 20,000) all weakly labeled (self-reported)
- Each sample in each dataset is a 12-lead ECG reading over x seconds with a 400hz sampling freq
  - Age and Gender as well

My Dataset: 3,200 samples (50/50 positive negative split)

DBN Implementation

Not many off-the-shelf packages exist.

I used pgmpy, because it had strong documentation and seemed the most credible

But it has drawbacks that I will get into later

Naive First Approach

DBN

- Latent variable: Chagas
- Dynamic Variables: 12-lead ECG Readings
- Static Age and Gender

At each time step we have:

- Edges from age and gender to Chagas
- Edges from Chagas to all 12 ECG readings
- Edges between Chagas across timesteps

Need to Down Sample and Discretize

Most samples have around 3,000 or 4,000 data points (every 1/400th of a second) for all 12-leads

- A DBN with 4,000 timesteps is not feasible and quite noisy
- Downsample to seconds and take the average of each of the 12 leads over the second
- **Problem: DBN only works with discrete data**
    - **All variables are continuous except gender**
    - **Assign bins to all continuous variables**

Naive Results (bad)

```
Accuracy: 0.5351681957186545
Confusion Matrix:
 [[161 166]
 [138 189]]
Precision: 0.532394366197183, Recall: 0.5779816513761468, F1 Score: 0.5542521994134897
```

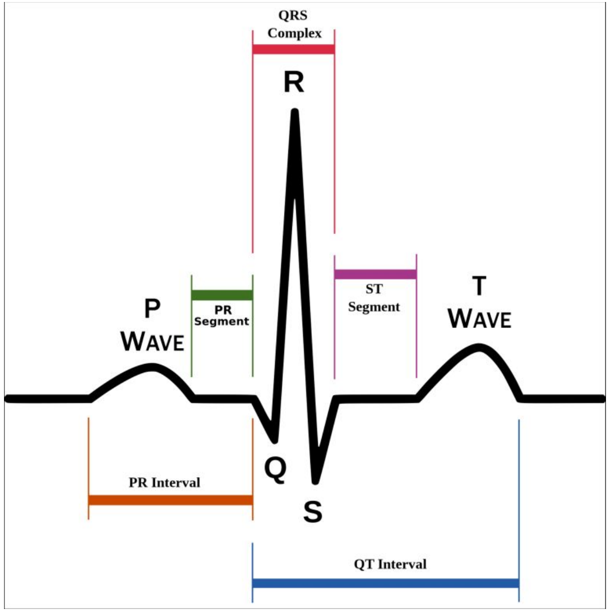
Wave Delineation Approach

In practice, doctors read ECG data to extract wave features

There is no closed form solution extracting these features

Computational Solutions Include:

1. Signal Processing
   a. simple
2. RNN/CNN
   a. Requires ECG data with wave feature labels



Wave Features

```
Features,Description,Data Type,Units
p_peaks,Amplitude of the P-wave,Float64,mV
p_onsets,Time at the onset of the P-wave,object,msec
p_offsets,Time at the offset of the P-wave,object,msec
q_peaks,Amplitude of the Q-wave,Float64,mV
r_onsets,Time at the onset of the R-wave,object,msec
r_offsets,Time at the offset of the R-wave,object,msec
s_peaks,Amplitude of the S-wave,Float64,mV
t_peaks,Amplitude of the T-wave,Float64,mV
t_onsets,Time at the onset of the T-wave,object,msec
t_offsets,Time at the offset of the T-wave,object,msec
sample_idx,Sample Idx,Int64,
lead,Lead,string,
heart_rate,Number of contractions of the heart per minute,Float64,bpm
r_peaks,Amplitude of the R-wave,object,mV
pr_interval,Time between onset of P-wave to onset of R-wave,Float64,msec
qrs_complex,Time between onset of Q-wave to offset of S-wave,Float64,msec
qt_interval,Time between onset of Q-wave to offset of T-wave,Float64,msec
rr_interval,Time between successive R-waves,Float64,msec
st_segment,Time between offset of S-wave to onset of T-wave,Float64,msec
```

DBN with beatwise timesteps

Each timestep is the onset or peak of the R wave (reliably detected)

- Feature are onsets/offsets of other waves
- If onset/offset occurs within +- x ms it counts as 1 else 0