Identification of Heart Disorders Using Symbolic Aggregate Approximation (SAX)

Moses Owusu

School of Public Health, Anschutz Medical Campus

Background

- Anomaly (irregular pattern) detection has gained much traction for its vast applications. Data is represented to reduce dimension but keep key information [5].
- Time series motifs are repeated patterns across a time series. Their similarity casts doubts on their occurrence being random.
- Motifs carry important information about the underlying dynamics just as Deoxyribonucleic acid (DNA) carries genetic information.
- These patterns appear with different frequencies, lengths, lags, and disparities across an entire series [2]. Time series discord refer to the most unusual time series subsequences.

Objectives

- Identify anomalies (discords) using SAX
- Classify & compare heart disorder ECG signals using LSTM and SAX

Methods

- Study data: Data was obtained from the MIT-BIH Arrhythmia database, consisting of 1000, 10-s (3600 non overlapping samples) ECG fragments (360 Hz) of 45 patients. It has 17 classes; 15 heart disorders a normal sinus pacemaker rhythm with at least 10 fragments for each.

	Classes	Fragments	Patients
Total	17	1000	45
Used	12	904	39
Males	58%	Females	42%
Age	32 - 89	years	

Table 1. Data description

	Filtering method		
Class	**Median**	**Low-pass**	**Wavelet**
APB	0.0648	0.1531	0.0003
AFIB	0.0527	0.1217	0.0003

Table 2. Filtering method & MSE

Data processing

SAX:

1. z-normalization.
2. PAA reduction
3. Find & implement SAX with optimal parameters
4. Identify discords & plot
5. Compute classification errors

LSTM:

1. Filtering (median, low-pass & wavelet)
2. Apply filter with lowest MSE
3. Perform feature selection
4. Implement LSTM
5. Compute accuracy

Piecewise Aggregate Approximation (PAA)

$$z_i = \frac{x_i - \mu}{\sigma} \qquad i = 1, 2, \ldots m$$

$$\bar{X}_i = \frac{m}{n}. \sum_{j=\frac{n}{m}(i-1)+1}^{(n/m).i} x_j$$

z_i = normalized x_i
μ = mean of X
σ = standard deviation of X
$X_n \approx \bar{X}_m$ where $m \leq n$
m = length of $\bar{X}$
n = length of **X**

PAA with different sizes

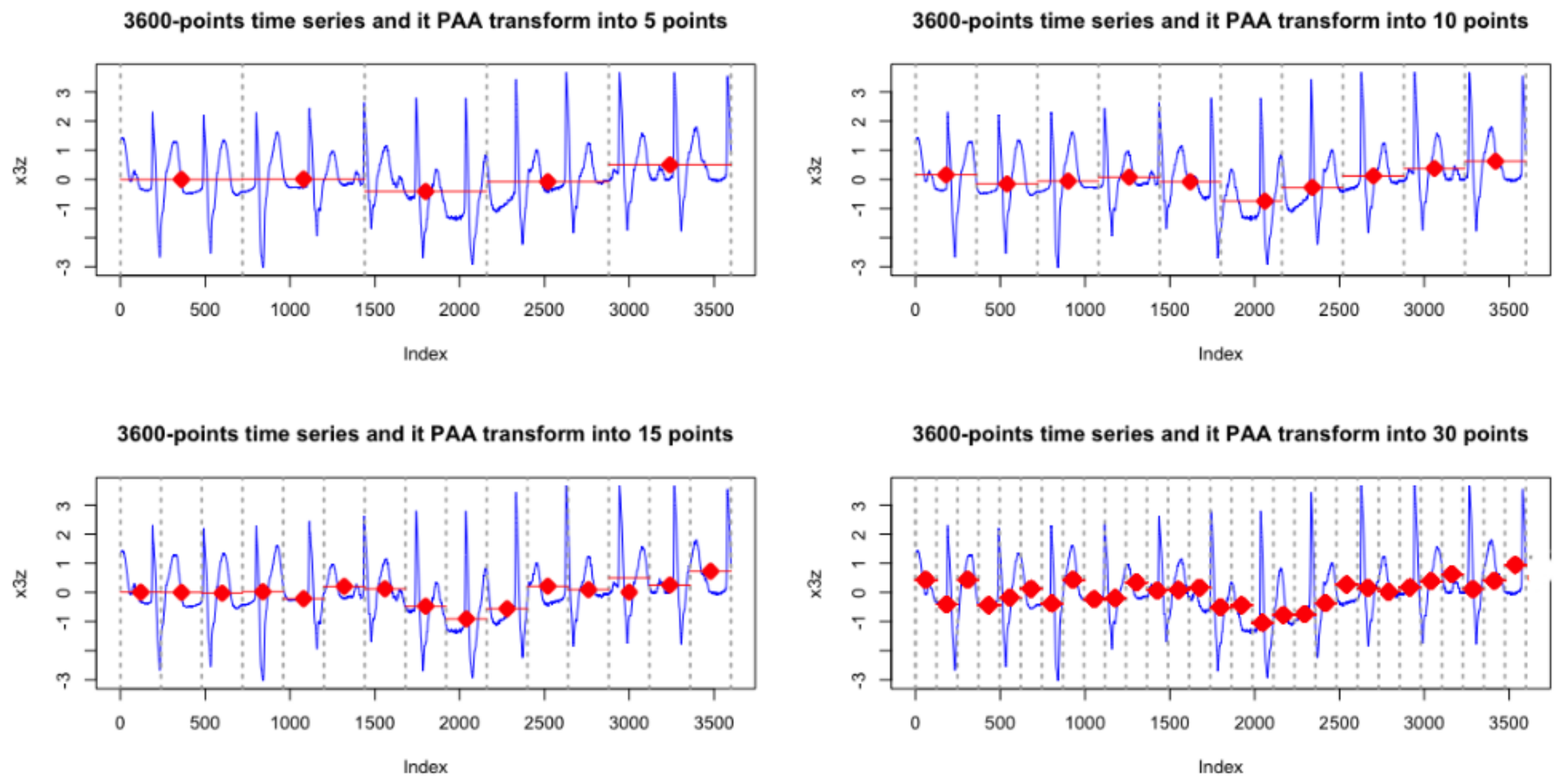


Figure 1. PAA with different parameters

Architecture of SAX and Lookup table

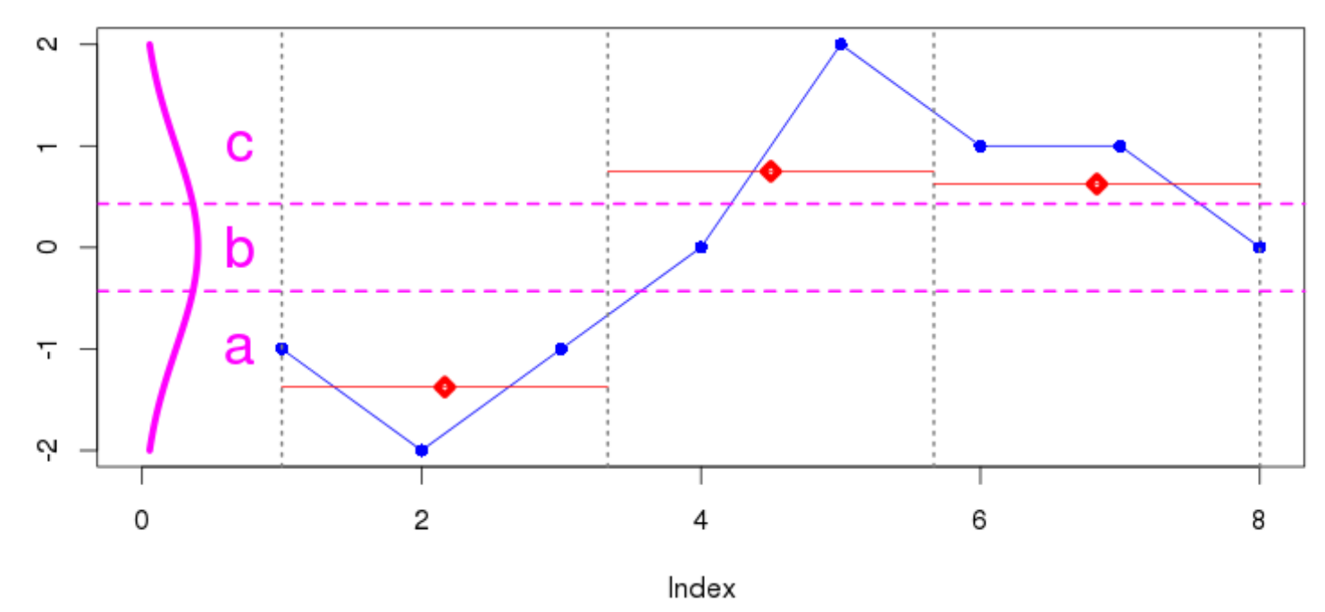


Figure 2. PAA & Character assignment [6]

$$MINDIST(\hat{Q}, \hat{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (dist(\hat{q}_i, \hat{c}_i))^2}$$

β / α	3	4	5
β_1	-0.43	-0.67	-0.84
β_2	0.43	0	-0.25
β_3	-	0.67	0.25
β_4	-	-	0.84

Table 3. Breakpoints Lookup table [3]

SAX word and sliding window

The breakpoints used to assign a,b,c in Figure 1 is Table 1.
The breakpoints are such that $\alpha = \alpha_1, \ldots, \alpha_{\beta-1}$ and for α_i to α_{i+1}, the area under the $N(0,1)$ curve is $\frac{1}{\alpha}$
MINDIST calculates SAX distance between words. For a word abc, "a" would be assigned to PAA terms between α_1 and α_2 with area $\frac{1}{3} = 0.33$, where β = 2 points.

$$Y_1 = \text{eeg}_1\text{eeg}_2\text{efg}_3\text{eff}_4\text{egh}_5\text{eeg}_6\text{efg}_7 \ldots \approx Y_2 = \text{eeg}_1\text{efg}_3\text{eff}_4\text{egh}_5\text{eeg}_6\text{efg}_7 \ldots$$

Y_1 reduced to Y_2 hence, reducing computational cost. Equation 1 then produces weights of words present in each class based on frequencies with which each word occurs.

Term Frequency - Inverse Document Frequency (TFIDF)

$$\text{tf}_{t,d} = \begin{cases} log(1 + f_{t,d}), & \text{if } f_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{idf}_{t,D} = \log \frac{|D|}{|d \in D : t \in d|} = \log \frac{N}{\text{df}_t}$$

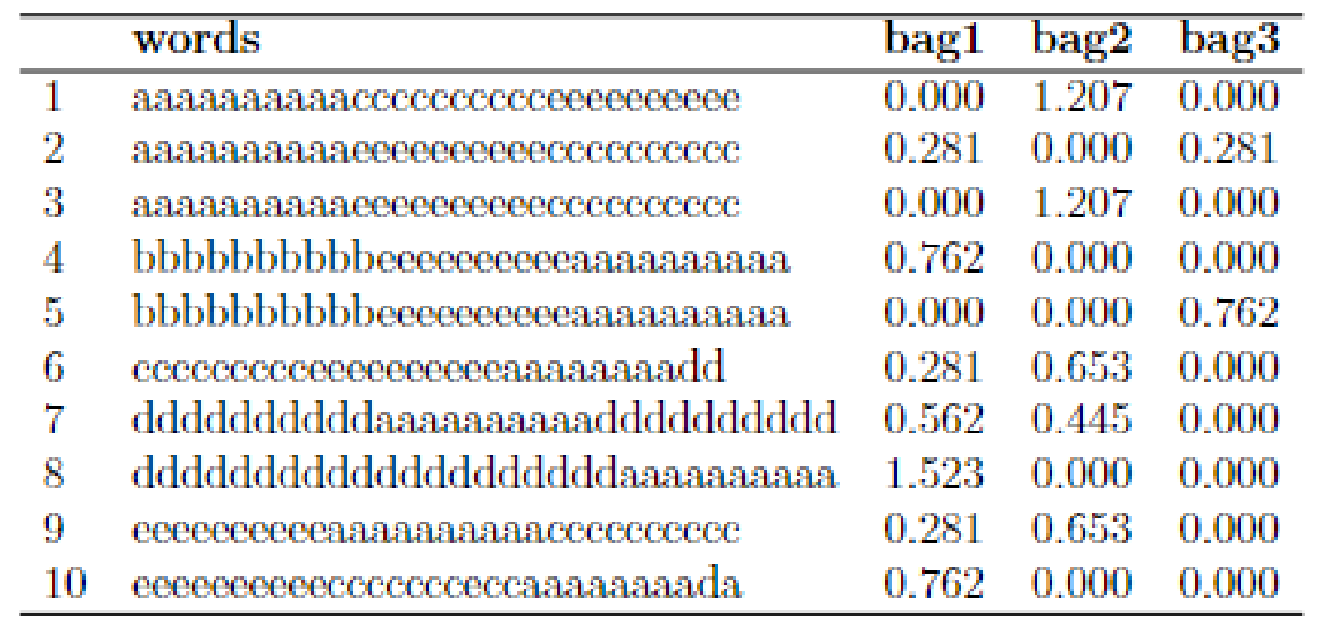
$f_{t,d}$ - frequency of t in d
tf- term frequency
idf - inverse tf
N - word bags cardinality
D - total number of classes
df_t - bags in which t occurs

t- term, d - word bag

the $tf * idf$ for t in the bag d of D set of bags is given as:

$$tf * idf(t, d, D) = tf_{t,d} \times idf_{t,D} tf * idf(t, d, D) = \log(1 + f_{t,d})$$

Results



	words	bag1	bag2	bag3
1	aaaaaaaaaacccccccccceeeeeeeeee	0.000	1.207	0.000
2	aaaaaaaaaaeeeeeeeeeecccccccccc	0.281	0.000	0.281
3	aaaaaaaaaaeeeeeeeeeecccccccccc	0.000	1.207	0.000
4	bbbbbbbbbbeeeeeeeeeeaaaaaaaaaa	0.762	0.000	0.000
5	bbbbbbbbbbeeeeeeeeeeaaaaaaaaaa	0.000	0.000	0.762
6	cccccccccceeeeeeeeeeaaaaaaaadd	0.281	0.653	0.000
7	ddddddddddaaaaaaaaaadddddddddd	0.562	0.445	0.000
8	ddddddddddddddddddddaaaaaaaaaa	1.523	0.000	0.000
9	eeeeeeeeeeaaaaaaaaaacccccccccc	0.281	0.653	0.000
10	eeeeeeeeeecccccccccccaaaaaaaada	0.762	0.000	0.000

Figure 3. TFIDF output

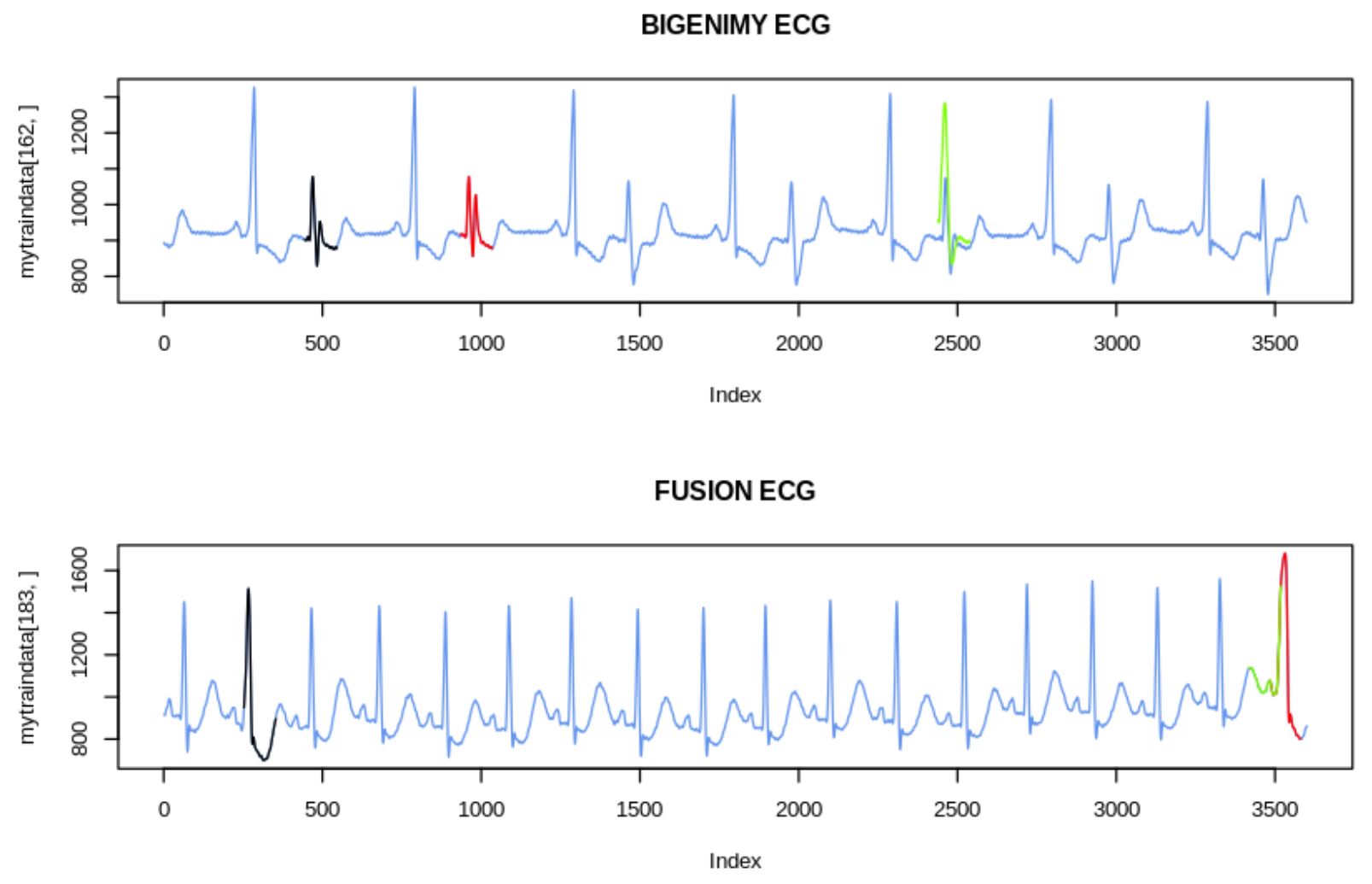


Figure 4. 3 best discords in each class

Class	n	Misclass	error	weighted
NSR	82	-	-	-
APB	45	45	0.3543	0.0079
SVTA	5	34	0.3864	0.0773
PVC	47	51	0.3953	0.0084
AFIB	30	30	0.2679	0.0089

Table 4. SAX classification error

TRIGENIMY and NSR WEIGHTED PATTERN PLOT

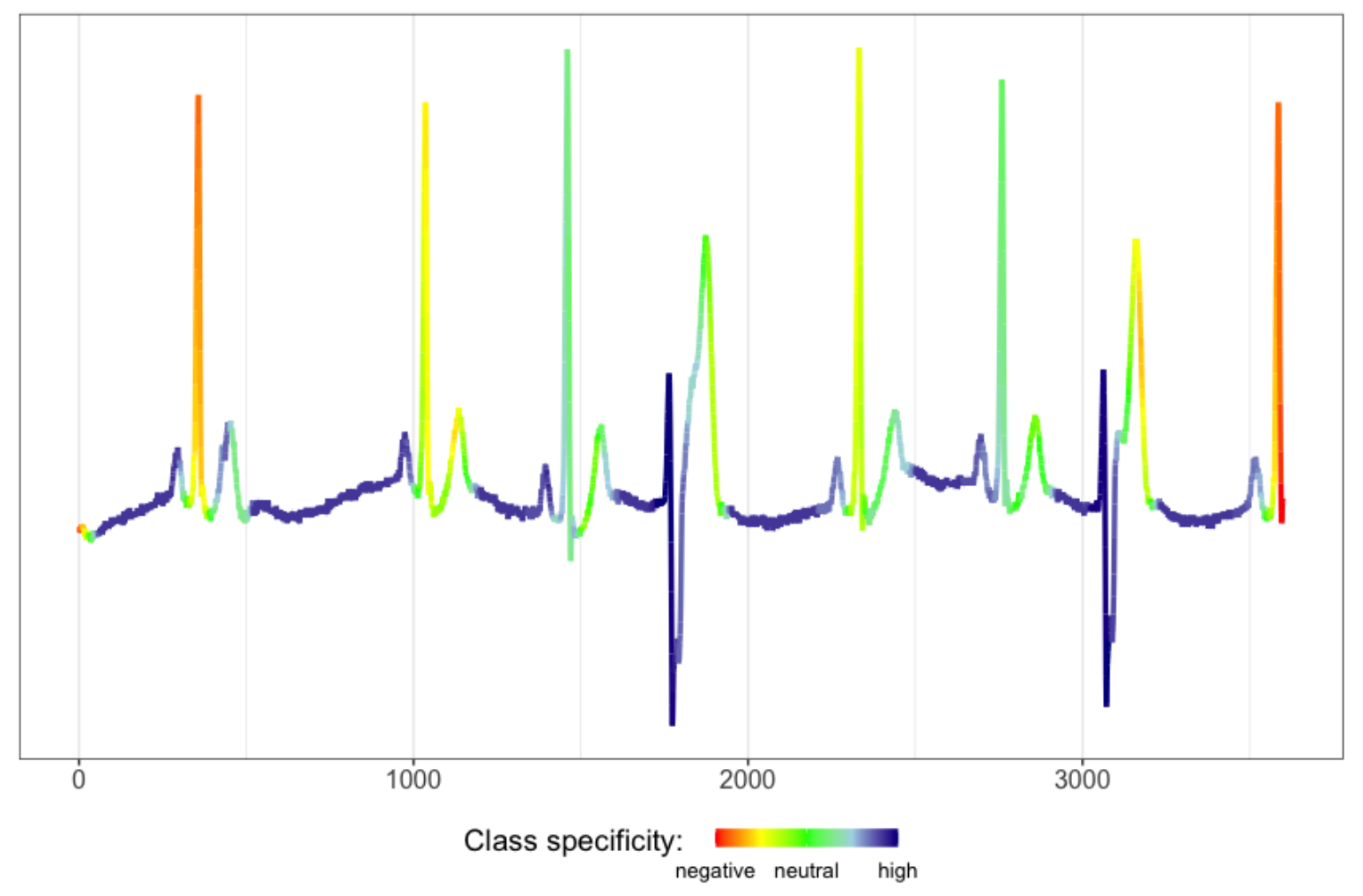


Figure 5. Weighted pattern plot

Conclusions and Limitations

- The results of SAX classification shows a good performance with low errors.
- Distance measure and position of best discord depict huge discrepancies between each disorder and NSR.
- This highlights differences in the ECG signals and suggest the presence of varying heart conditions as was initially planned.
- Similar location of best discords and equal distance measures would have meant all the ECG are the same.
- LSTM does not factor in position of best discord in its computation so lower accuracies were obtained.
- The challenge of these algorithms is the non-existence of prior knowledge about the precise positions of these patterns.
- The use of exhaustive search approaches lead to very high computational cost
- A small subsequence length leads to identifying too many patterns which may not necessarily be significant. Longer length sequences may also miss key patterns.

References

[1] Lena Biel, Ola Pettersson, Lennart Philipson, and Peter Wide.
Ecg analysis: a new approach in human identification.
IEEE transactions on instrumentation and measurement, 50(3):808–812, 2001.

[2] Tian Huang, Yongxin Zhu, Yafei Wu, and Weiwei Shi.
J-distance discord: an improved time series discord definition and discovery method.
In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 303–310. IEEE, 2015.

[3] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi.
Experiencing sax: a novel symbolic representation of time series.
Data Mining and knowledge discovery, 15:107–144, 2007.

[4] Abdullah Mueen.
Time series motif discovery: dimensions and applications.
Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(2):152–159, 2014.