

# Multivariate Symbolic Aggregate Approximation for ECG Analysis

Moritz M. Konarski

Supervised by Prof. Taalaibek M. Imanaliev

May 3, 2021

## Outline

## Contents

1	Introduction	1
2	Methods	5
3	Results	9

## 1 Introduction

### ECG Basics

#### What is an ECG?

- electrocardiogram (ECG or EKG) records the heart's electrical activity
- contains up to 12 simultaneous measurements—the leads
- common medical diagnostic tool

### ECGs as Time Series

**Definition 1.** A discrete time series is an ordered sequence that, at discrete points in time, has  $n$  values each. If  $n = 1$ , the series is univariate and if  $n > 1$ , it is multivariate.

- digital ECGs are discrete multivariate time series:
  - have  $> 1$  value at each point, often  $n = 12$
  - recorded at discrete, evenly spaced time points
- time series analysis methods can be applied to ECGs

Lead MLII of MIT-BIH/103

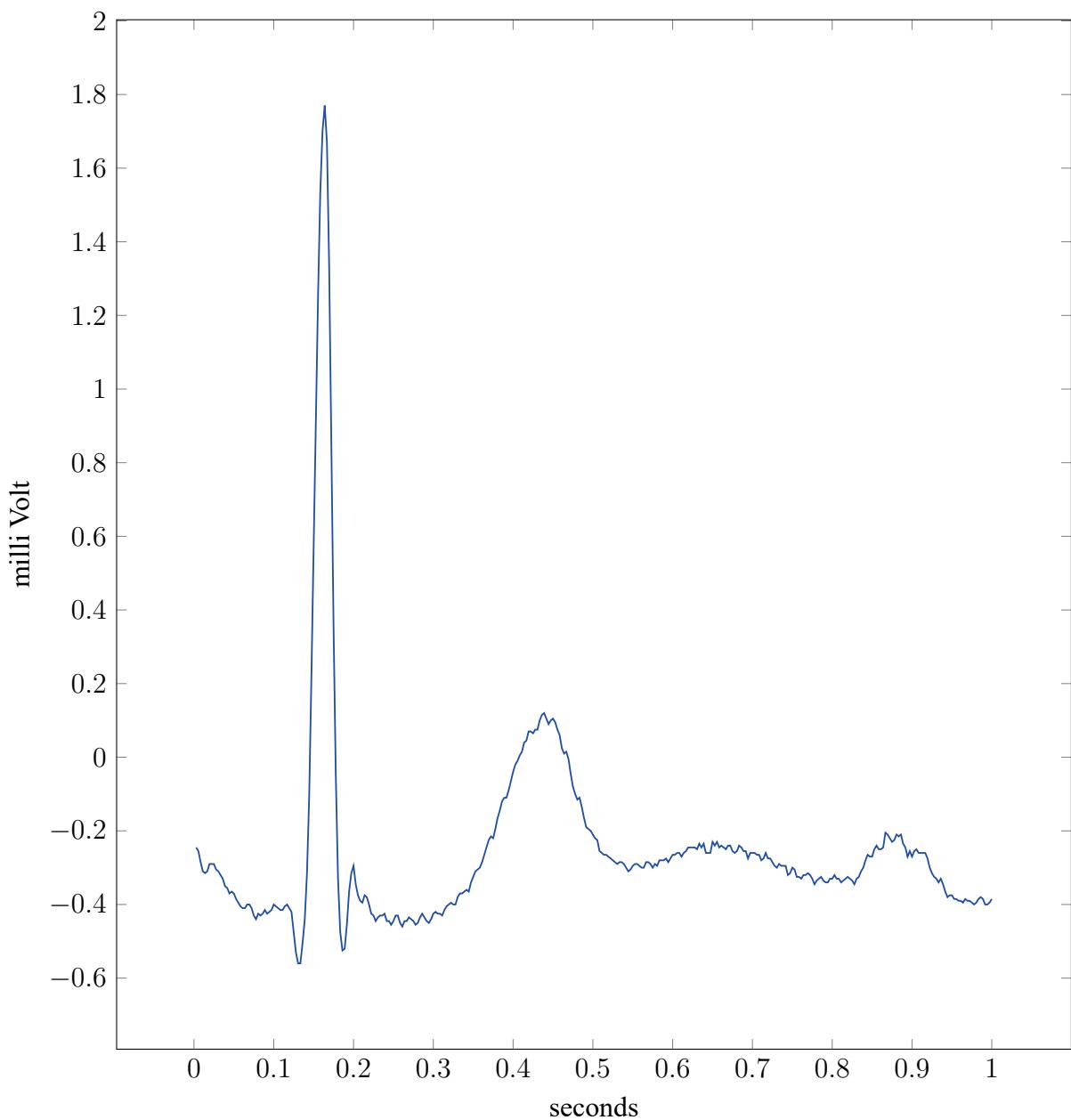


Figure 1: ECG of one heartbeat

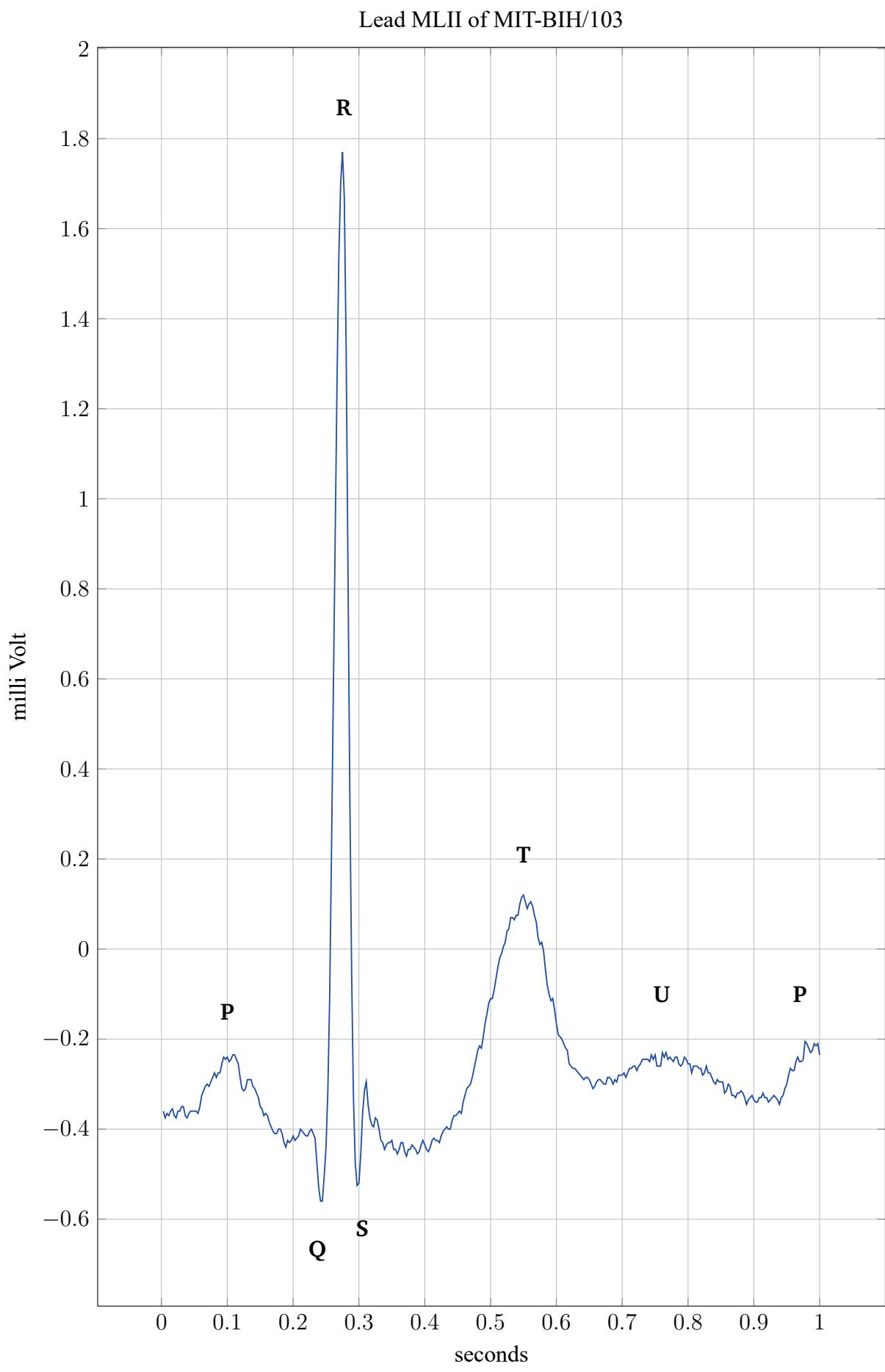


Figure 2: Annotated ECG of one heartbeat

## ECG Analysis

### ECG Analysis

- standard method: manual analysis by cardiologist
- recently: automated or computer-assisted ECG analysis
- multiple stages: (1) signal acquisition; (2) data transformation, processing, filtering; (3) waveform recognition, feature extraction; (4) classification
- current research focus: artificial neural networks
- relatively new methods are SAX, MSAX, and HOTSAX

### SAX, MSAX, and HOTSAX

- Lin *et al.* (2003): Symbolic Aggregate Approximation (SAX)—simplified, symbolic representation
- Anacleto *et al.* (2020): Multivariate SAX (MSAX)—expands SAX to multivariate time series
- Keogh *et al.* (2005): Heuristically Ordered Time series using Symbolic Aggregate Approximation (HOTSAX)—discord discovery algorithm for SAX

### Time Series Discords

**Definition 2.** A time series discord is the subsequence of a time series that is most different from all other subsequences.  $k$  time series discords are the  $k$  most different subsequences.

- discords represent anomalies in an ECG
- can be found by comparing all subsequences to all other subsequences; does not scale well
- HOTSAX makes this process faster

## Hypothesis

### Hypothesis

HOTSAX with MSAX will increase the number of relevant discords detected compared to HOTSAX with SAX. Accuracy can be judged with the help of annotated ECGs from online databases.

## 2 Methods

### SAX and MSAX

#### Step 1: Z-Normalization

##### Assumption

The time series values are normally distributed.

SAX

- normalize univariate time series
- uses scalar mean and variance

MSAX

- normalize multivariate time series
- uses vector mean and covariance matrix

#### Step 2: Dimensionality Reduction

##### PAA

Piecewise Aggregate Approximation (PAA) takes  $T$  time series points, splits it into  $w$  ( $w < T$ ) segments, and averages each of them.

SAX

- apply PAA to time series

MSAX

- apply PAA to each of the time series individually

#### Step 3: Discretization

##### SAX Discretization

Find breakpoints splitting  $\mathcal{N}(0, 1)$  into  $B$  equiprobable segments. Assign a letter to each area:  $a$  to most-negative,  $b$  to the next biggest... PAA segments get letters based on which area they are in.

SAX

- discretize the time series
- results in one word

MSAX

- discretize each time series individually
- results in one word, one letter per time series

# SAX PAA of lead ML

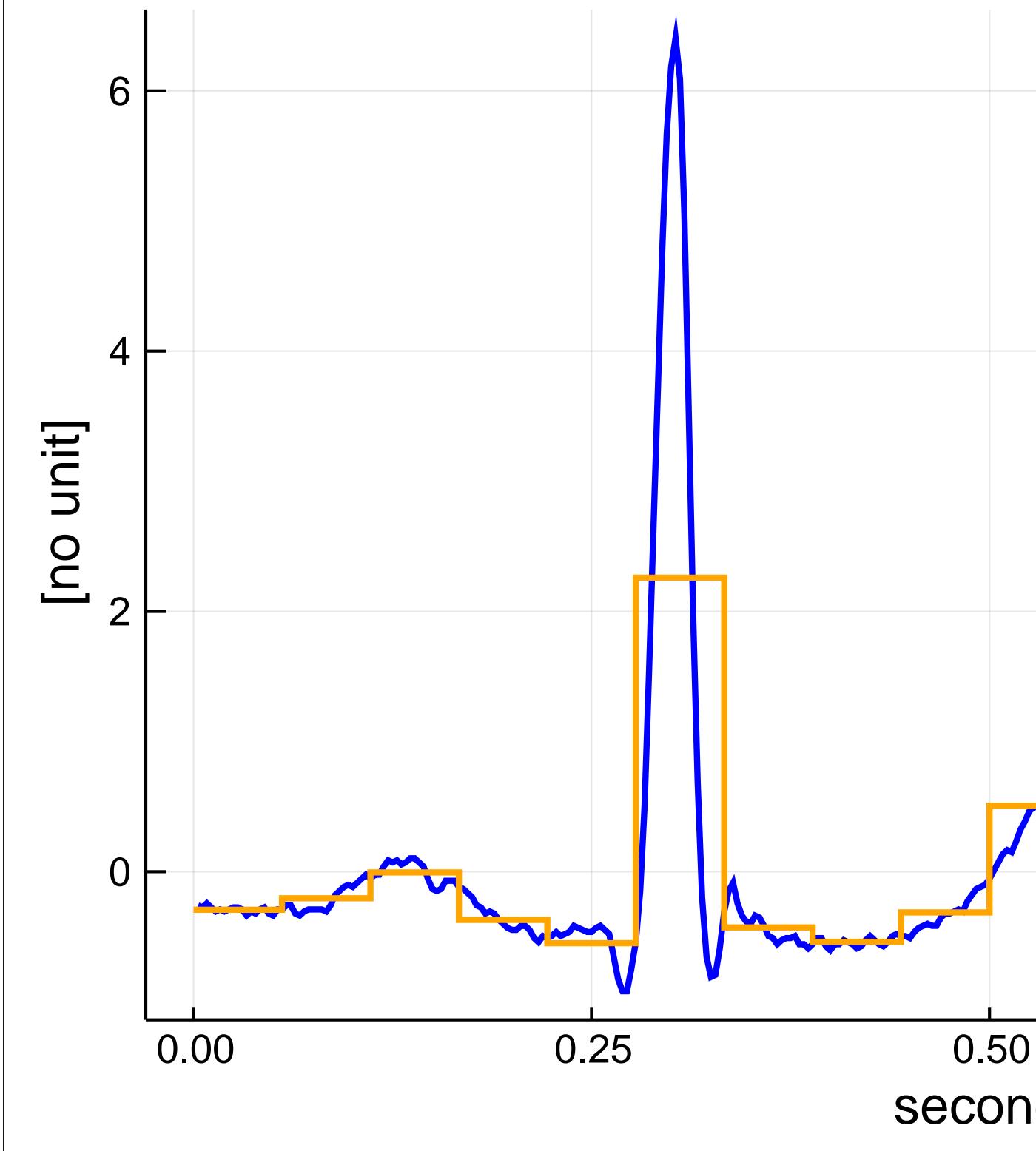


Figure 3: ECG with PAA (MITBIH/100,  $w = 18$ ,  $T = 360$ )

# SAX of lead MLII c

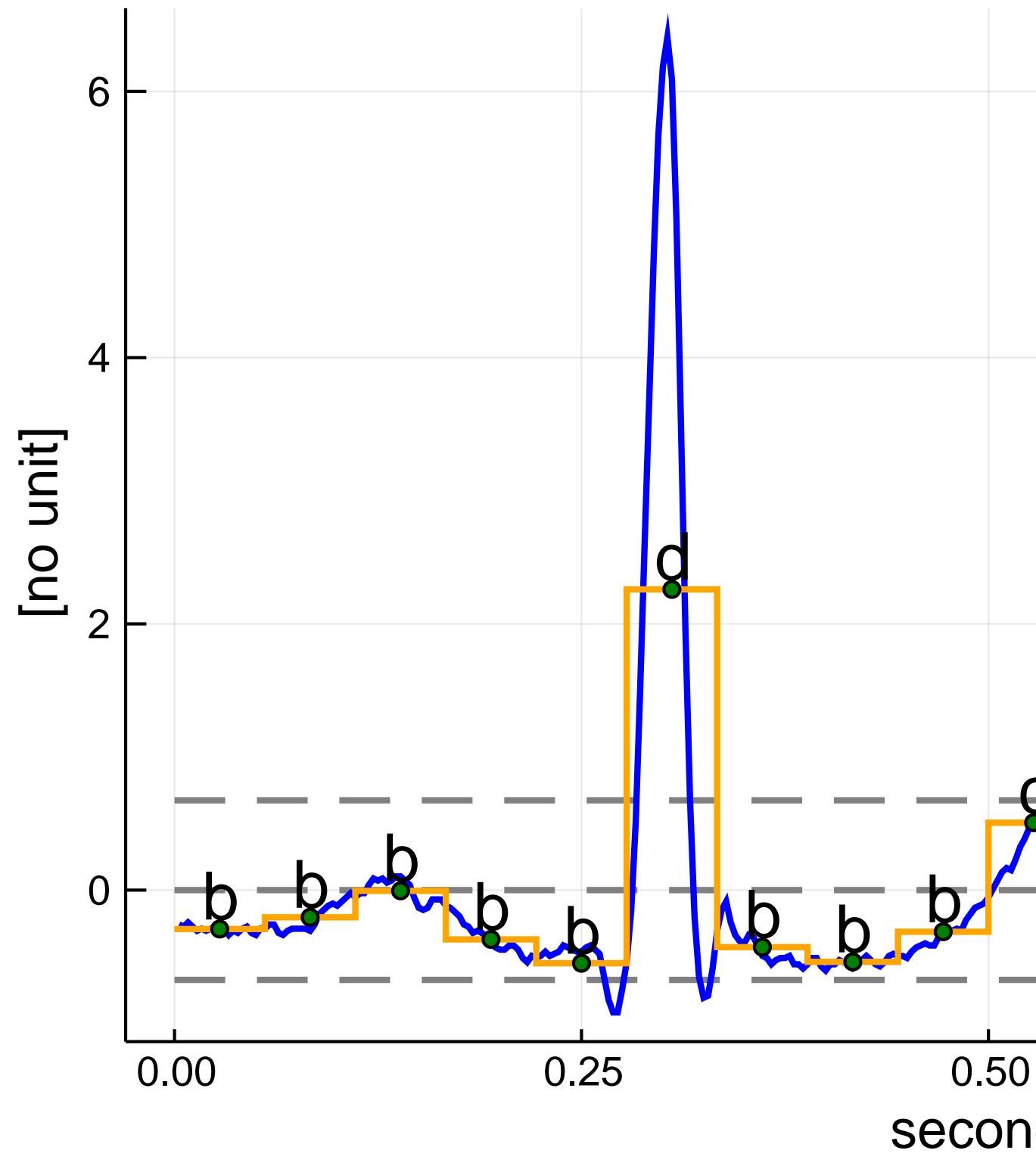


Figure 4: SAX (MITBIH/100,  $w = 18, T = 360, B = 4$ )

Table 1: Difference matrix for  $B = 4$

	a	b	c	d
a	0	0	0.67449	1.34898
b	0	0	0	0.67449
c	0.67449	0	0	0
d	1.34898	0.67449	0	0

#### Step 4: Distance Measure

##### MINDIST

A distance measure is defined to compare two SAX words. Distance is defined for a pair of letters: 0 if they are neighbors; absolute difference of breakpoint values otherwise.

SAX

$$\sqrt{\frac{T}{w}} \sqrt{\sum_{i=1}^w (\text{dist}(\hat{q}[i], \hat{c}[i]))^2}$$

MSAX

$$\sqrt{\frac{T}{w}} \sqrt{\sum_{i=1}^w \left( \sum_{j=1}^n (\text{dist}(\hat{q}_j[i], \hat{c}_j[i]))^2 \right)}$$

#### Difference Matrix

#### HOTSAX

##### HOTSAX

- “brute-force” discord discovery is slow, needs  $T^2$  operations
- HOTSAX speeds up discord discovery by considering that
  - discords are rare, start with rarest segment
  - similar segments have similar distances, consider together
- HOTSAX detects anomalies, it is not a classifier
- it uses SAX and MSAX for dimensionality reduction

### **3 Preliminary Results**

#### **Implementation**

##### **Implementation**

- SAX, MSAX, HOT SAX implemented in Julia, a scientific programming language
- used annotated digital ECGs from the MIT-BIH arrhythmia database
- HOT SAX performed for different  $w$ ,  $B$ , subsequence lengths
- results exported to CSV file and analyzed using the R programming language

#### **Preliminary Results**

##### **Preliminary Results**

- focus on comparing SAX and MSAX with the top  $k = 80$  discords
- to analyze the relevance of results, recall (sensitivity) is used
- analyzed total of 816 results for different parameters (SAX and MSAX for each)
- recall for MSAX is higher compared to SAX
- if SAX is applied to 2 leads and the results combined, it slightly outperforms MSAX

#### **Outlook**

##### **Outlook**

- perform statistical tests for significance of the result
- analyze the outliers visible in the boxplots
- more tests with different sets of parameters
- explore the influence of parameters on the result
- use the 12-lead INCART ECG database to investigate the influence of larger numbers of leads

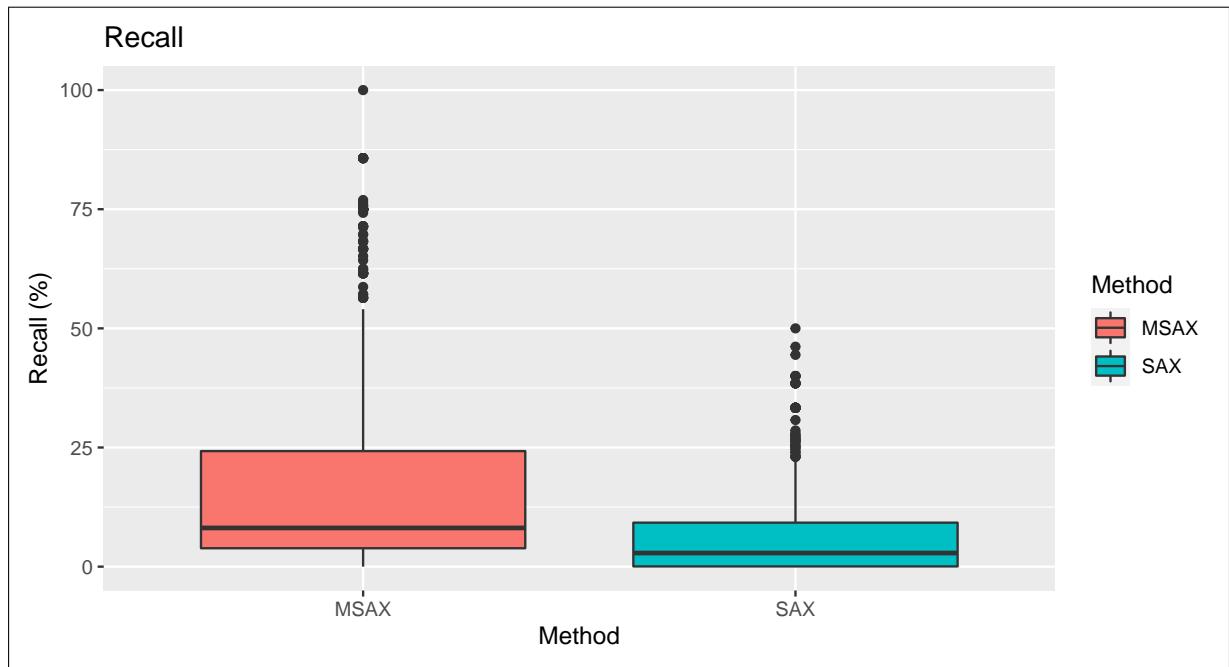


Figure 5: Boxplot comparing Recall for MSAX and single-lead SAX

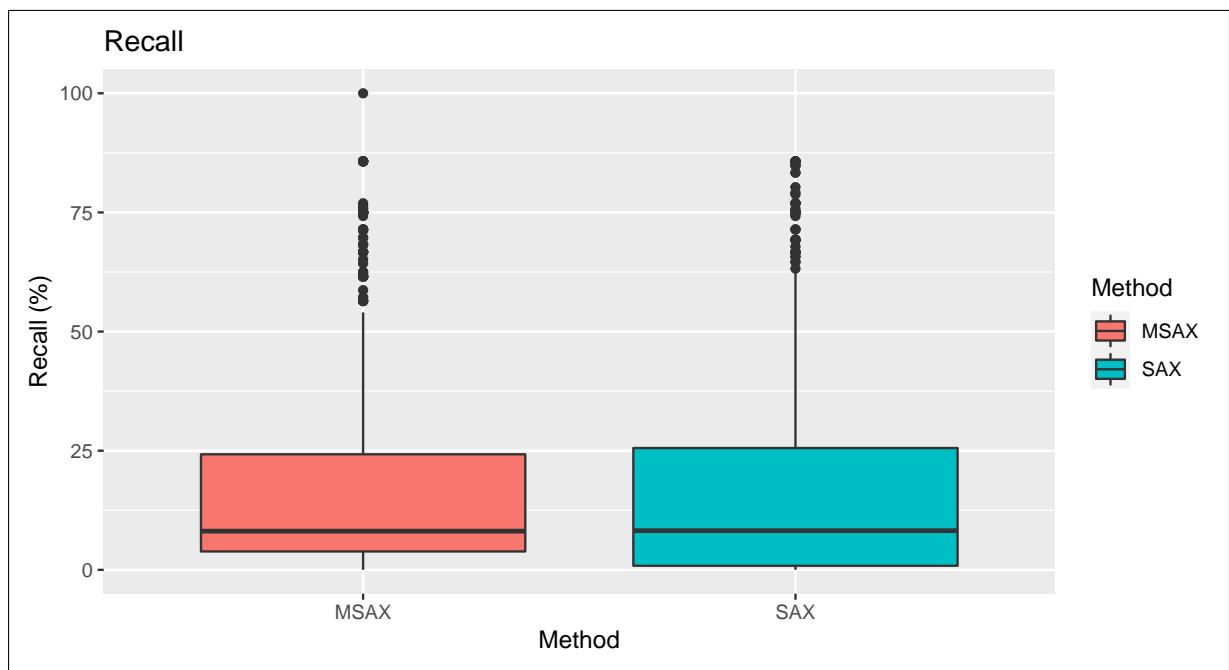


Figure 6: Boxplot comparing Recall for MSAX and dual-lead SAX

# Thank You

## Q & A

Thank You!

# References

## References

# References

- [1] G. B. Moody and R. G. Mark, *MIT-BIH Arrhythmia Database*, physionet.org, 1992. doi: 10.13026/C2F305.
- [2] “The top 10 causes of death.” (), [Online]. Available: <https://www.who.int/news-room/factsheets/detail/the-top-10-causes-of-death> (visited on 05/02/2021).
- [3] M. Anacleto, S. Vinga, and A. M. Carvalho, “MSAX: Multivariate Symbolic Aggregate Approximation for Time Series Classification,” in *Computational Intelligence Methods for Bioinformatics and Biostatistics*, P. Cazzaniga, D. Besozzi, I. Merelli, and L. Manzoni, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 90–97. doi: 10.1007/978-3-030-63061-4\_9.
- [4] Kligfield Paul *et al.*, “Recommendations for the Standardization and Interpretation of the Electrocardiogram,” *Circulation*, vol. 115, no. 10, pp. 1306–1324, 2007. doi: 10.1161/CIRCULATIONAHA.106.180200.
- [5] L. Xie *et al.*, “Computational Diagnostic Techniques for Electrocardiogram Signal Analysis,” *Sensors*, vol. 20, no. 21, p. 6318, Nov. 5, 2020. doi: 10.3390/s20216318.
- [6] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03*, San Diego, California: ACM Press, 2003, pp. 2–11. doi: 10.1145/882082.882086.
- [7] C. Zhang *et al.*, “Anomaly detection in ECG based on trend symbolic aggregate approximation,” *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2154–2167, 2019, ISSN: 1547-1063. doi: 10.3934/mbe.2019105.
- [8] E. Keogh, J. Lin, and A. Fu, “HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence,” in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, USA: IEEE, 2005, pp. 226–233. doi: 10.1109/ICDM.2005.79.