Multivariate Symbolic Aggregate Approximation for ECG Analysis

Moritz M. Konarski
Supervised by Prof. Taalaibek M. Imanaliev

Applied Mathematics and Informatics Program,
American University of Central Asia

May 3, 2021
Bishkek, Kyrgyz Republic

Outline

Introduction

What is an ECG?

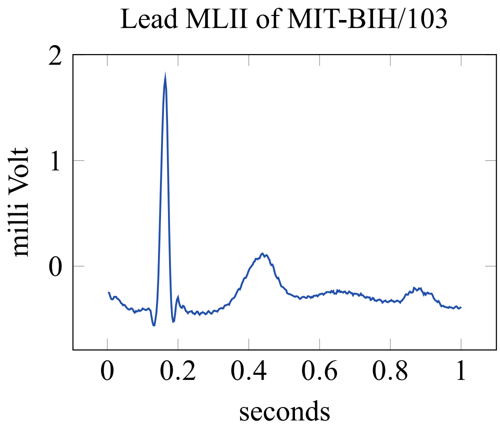


Figure 1: ECG of one heartbeat

- electrocardiogram (ECG or EKG) records the heart's electrical activity
- contains up to 12 simultaneous measurements—the leads [1]
- common medical diagnostic tool[2]

What is an ECG?

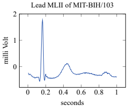


Figure 1: ECG of one heartbeat

- electrocardiogram (ECG or EKG) records the heart's electrical activity
- contains up to 12 simultaneous measurements—the leads [1]
- common medical diagnostic tool[2]

- muscle contractions caused by electric pulses
- electric pulse can be measured on the skin
- the measuring things are called electrodes
- electrodes form leads (need 2 to measure anything)
- they have specific positions and names
- 12 leads is the modern standard
- most types of heart disease can be detected
- diagnosis and analysis is performed by trained cardiologists
- **datasets available online; contain 2 or more leads (the most significant ones)**
- **I will be using online datasets for my analysis**
- heart diseases are some of the most deadly ones, thus ECG are really important

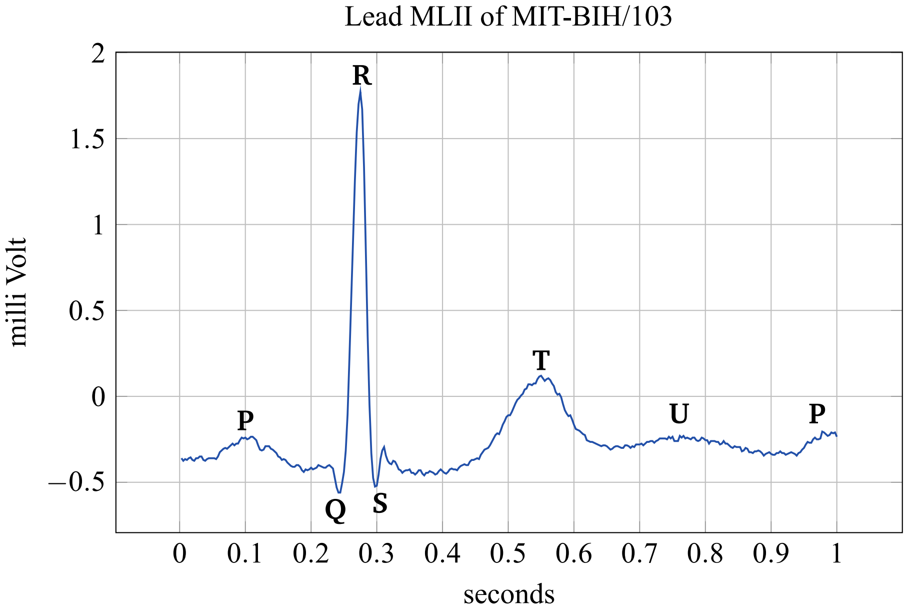


Figure 2: Annotated ECG of one heartbeat

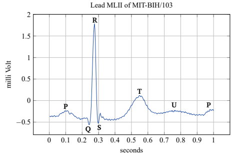


Figure 2: Annotated ECG of one heartbeat

- P wave: atria depolarizing / blood entering the heart
- QRS complex: ventricular depolarization / heart contraction pumping blood
- T: return of ventricle to polarized state
- U: present in 25%, maybe some feedback
- P wave: atria depolarizing / blood entering the heart
- ST-segment: significant, depression, elevation, slope show ischaemia
- R-R interval: shows rhythm and thus arrhytmia etc

ECGs as Time Series

Definition

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: [3]
  - 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

ECGs as Time Series

Definition

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: [3]
  - 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

- modern ECGs have at least 2, most have 12
- digital ones have set sampling frequencies, even the machines have set frequencies
- multivariate: measure more than 1 lead per time point
- discrete: set sample frequency in the machines
- discrete: because measured at discrete moments in time
- time series: they are data measured at equal time intervals
- n measurements per point in time (i.e. leads)
- $n = 1$ is univariate, $n > 1$ is multivariate

ECG Analysis

- standard method: manual analysis by cardiologist
- recently: automated or computer-assisted ECG analysis
- multiple stages: [4](1) signal acquisition; (2) data transformation, processing, filtering; (3) waveform recognition, feature extraction; (4) classification
- current research focus: artificial neural networks…[5]
- relatively new methods are SAX, MSAX, and HOTSAX

ECG Analysis

- standard method: manual analysis by cardiologist
- recently: automated or computer-assisted ECG analysis
- multiple stages: [4](1) signal acquisition; (2) data transformation, processing, filtering; (3) waveform recognition, feature extraction; (4) classification
- current research focus: artificial neural networks…[5]
- relatively new methods are SAX, MSAX, and HOTSAX

- is relatively slow; time is of the essence
- lots of training required
- error prone
- maybe not feasible for long ECGs
- can speed up process
- can pick up details humans miss
- digitizing paper ECGs or recording digital ones
- filtering to remove various types of noise
- reduce complexity of the data
- select important features and neglect irrelevant ones to ease analysis
- often added, figure out if there is some disease present or not
- balance between accuracy and complexity needed
- ann: hand all the steps discussed to a NN; use as good classifier too

SAX, MSAX, and HOTSAX

- Lin *et al.* (2003): Symbolic Aggregate Approximation (SAX)—simplified, symbolic representation [6] [7]
- Anacleto *et al.* (2020): Multivariate SAX (MSAX)—expands SAX to multivariate time series [3]
- Keogh *et al.* (2005): Heuristically Ordered Time series using Symbolic Aggregate Approximation (HOTSAX)—discord discovery algorithm for SAX

SAX, MSAX, and HOTSAX

- Lin *et al.* (2003): Symbolic Aggregate Approximation (SAX)—simplified, symbolic representation [6] [7]
- Anacleto *et al.* (2020): Multivariate SAX (MSAX)—expands SAX to multivariate time series [3]
- Keogh *et al.* (2005): Heuristically Ordered Time series using Symbolic Aggregate Approximation (HOTSAX)—discord discovery algorithm for SAX

- ecg as letters that mean same thing as original
- guaranteed to behave like the original data
- works on univariate time series
- has been used on ECGs
- takes the correlation between ecg leads into account
- cov mat: covariance between each lead and variance on diag
- uses sax representation to make the finding of discords easier
- can use MSAX just as well

Time Series Discords

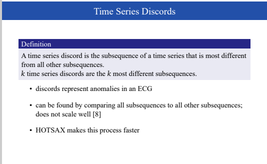
Definition

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 [8]
- HOTSAX makes this process faster

2021-05-03

MSAX for ECG Analysis
└─Introduction
└─ECG Analysis
└─Time Series Discords



- these can be diseases, noise, etc
- the discord does not discern
- this is not feasible because of complexity

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.

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.

- mention that MSAX to ECGs in particular is new
- mention that HOTSAX with MSAX is new
- results will not be great as HOTSAX is not a real classifier; this is about finding out if MSAX adds more useful information
- THIS METHOD WILL NOT BE SUPER ACCURATE; MANY ECG changes are relatively small and would get lost in the SAX process
- THE METHOD HAS NO AWARENESS OF MEDICAL RELEVANCE OR ANY OF THAT
- NOVEL: MSAX IS SUPER NEW; HAS NOT BEEN APPLIED TO ECGs IN THIS WAY; ALSO NOT USED WITH HOTSAX

Methods

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 1: Z-Normalization

Assumption

The time series values are normally distributed.

SAX	MSAX
• normalize univariate time series	• normalize multivariate time series
• uses scalar mean and variance	• uses vector mean and covariance matrix

- say that the process is the same as MSAX based on SAX
- this is assumed and this worked for other people who applied SAX to ECGs
- to compare time series, normalization is the accepted step
- what is this
- takes into account the correlation between leads

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 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	MSAX
• apply PAA to time series	• apply PAA to each of the time series individually

- this reduces complexity
- PAA form of time series is shorter and simpler
- it still somewhat corresponds to the original

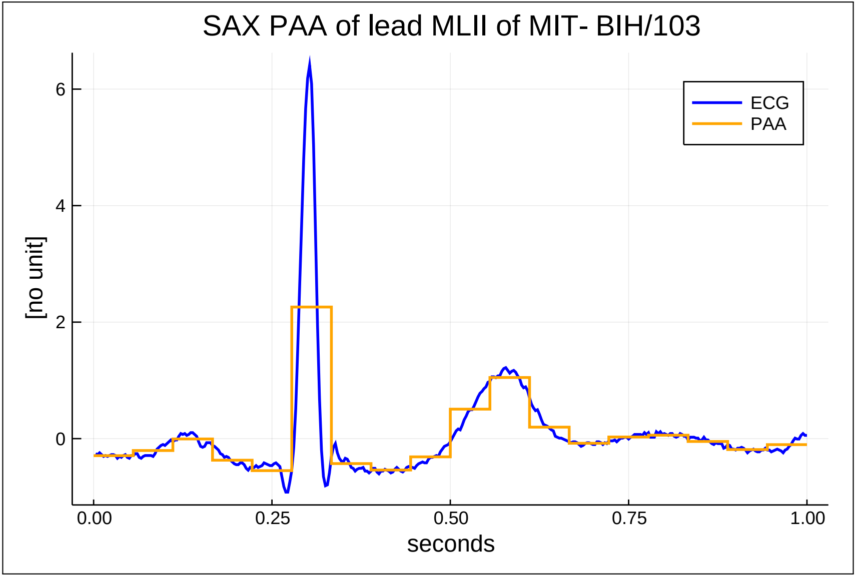


Figure 3: ECG with PAA (MITBIH/100, $w = 18$, $T = 360$)

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

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

- result is called word
- N is the alphabet size
- big thing here is that this gives defined probability to each letter; makes no sense for real numbers (like PAA values)
- simplifies time series even more
- creates discrete categories, can be more useful

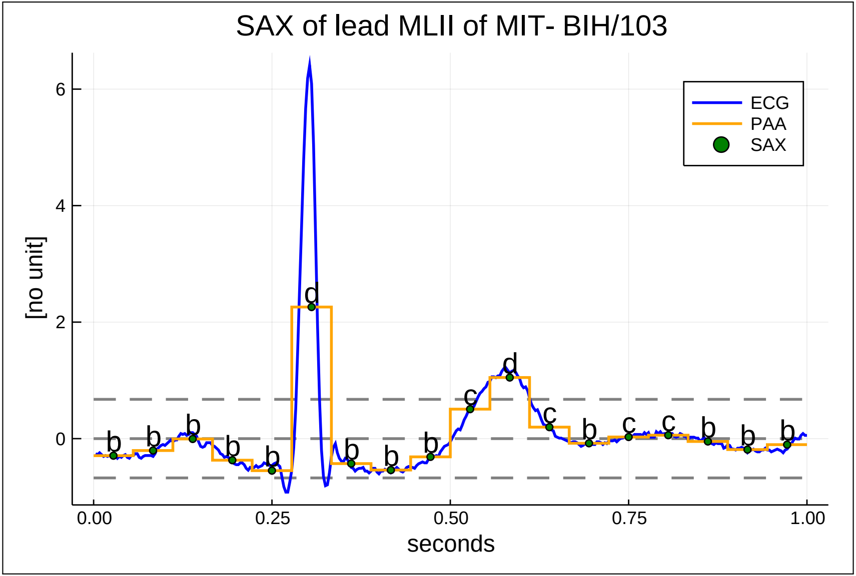


Figure 4: SAX (MITBIH/100, $w = 18$, $T = 360$, $B = 4$)

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

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

MSAX

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

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

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

MSAX

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

- n – length of original time series
- w – length of word
- this lower-bounds the euclidean distance, meaning that results in SAX should hold true for the real data too

Difference Matrix

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

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

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

- this is the basic idea that can speed up the process
- it is not guaranteed to do so, but it does not decrease efficiency
- this speeds up the process even more as we have fewer elements
- because of lower bounding, it still gives accurate results

Results

Implementation

- SAX, MSAX, HOTSAX implemented in Julia, a scientific programming language
- used annotated digital ECGs from the MIT-BIH arrhythmia database [1]
- HOTSAX performed for different w, B, subsequence lengths
- results exported to CSV file and analyzed using the R programming language

Implementation

- SAX, MSAX, HOTSAX implemented in Julia, a scientific programming language
- used annotated digital ECGs from the MIT-BIH arrhythmia database [1]
- HOTSAX performed for different w, B, subsequence lengths
- results exported to CSV file and analyzed using the R programming language

- fast, type support, great libraries, JIT compilation
- ecgs have all heart beats annotated
- know which are normal, diseases, noise, etc
- 48 recordings of 30 minutes
- w - paa segments; B - alphabet size; subsequence length for HOTSAX

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

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

- how many relevant items are selected
- recall = true positive / (true positive + false negative)
- this is done because for medical things it is more useful to look at a couple too many segments than not enough

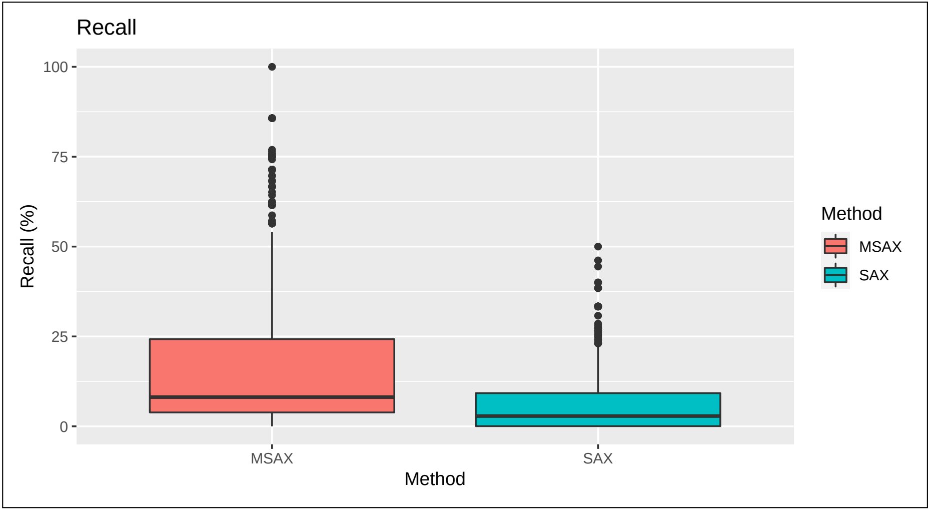


Figure 5: Boxplot comparing Recall for MSAX and single-lead SAX

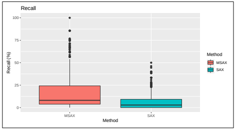


Figure 5: Boxplot comparing Recall for MSAX and single-lead SAX

- msax: average = 17.5%
- sax: average = 6.4%

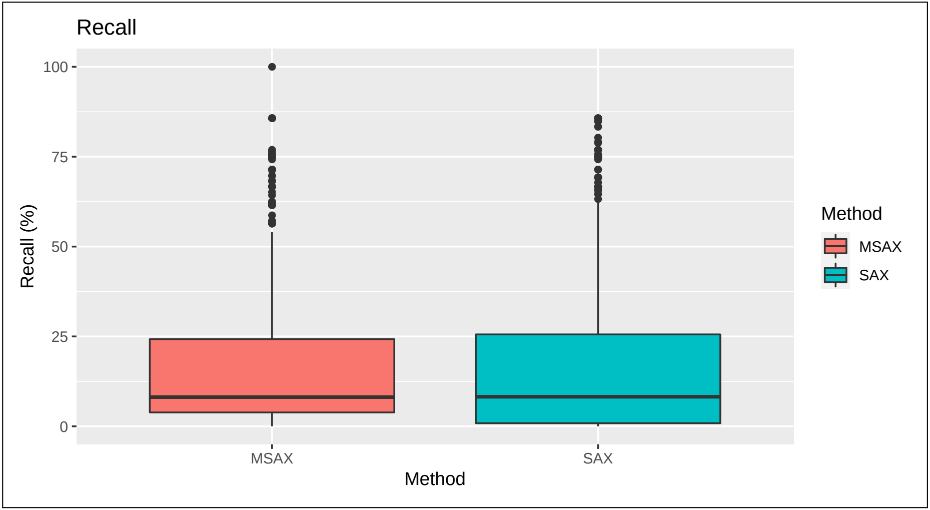


Figure 6: Boxplot comparing Recall for MSAX and dual-lead SAX

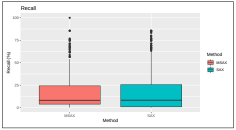


Figure 6: Boxplot comparing Recall for MSAX and dual-lead SAX

- msax: average = 17.5%
- sax: average = 18.5%

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

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

- for example t-test, biserieal correlation

Thank You!

References I

[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/fact-sheets/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, Gettes Leonard S., Bailey James J., *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.

References II

[5] L. Xie, Z. Li, Y. Zhou, *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, Y. Chen, A. Yin, *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.