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

Partially supported by AUCA FRG
“Mathematical Model in Acute Cardiac Ischemia Evaluation”

May 31, 2021
Bishkek, Kyrgyz Republic

Outline

1. Introduction and Background
2. Methods
3. Results
4. Conclusion

Introduction

- ischemic heart disease causes 16% of global deaths, can be diagnosed using electrocardiogram (ECG or EKG)
- an ECG records the heart's electrical activity
- manual (by cardiologist) ECG analysis is slow and error-prone → computers can help
- long ECGs are problematic even for computers → representation needed
- representations are simpler and still correspond to raw data

Representation and Classification

- time series can be represented using SAX, MSAX
- representation can be analyzed instead of raw data
- HOT SAX can be used to classify ECG segments as “discord”, “non-discord”
- this work combines MSAX and HOT SAX into HOT MSAX
- effectiveness of methods judged by recall and precision

Research Questions & Hypothesis

- Using the MIT-BIH ECG database, what parameters maximize HOT SAX and HOT MSAX recall?
- Which is better: optimal HOT SAX or optimal HOT MSAX?

HOT MSAX should have higher recall than HOT SAX if both use their best parameters

SAX and MSAX – Overview

SAX	MSAX
Application	
univariate time series, e.g. a single ECG lead	multivariate time series, e.g. multiple ECG leads (whole ECG)
Steps	
(1) univariate z-normalization (2) PAA dimension reduction (w) (3) SAX discretization (a)	(1) multivariate z-normalization (2) PAA dimension reduction (w) (3) SAX discretization (a)

SAX and MSAX – Step (2)

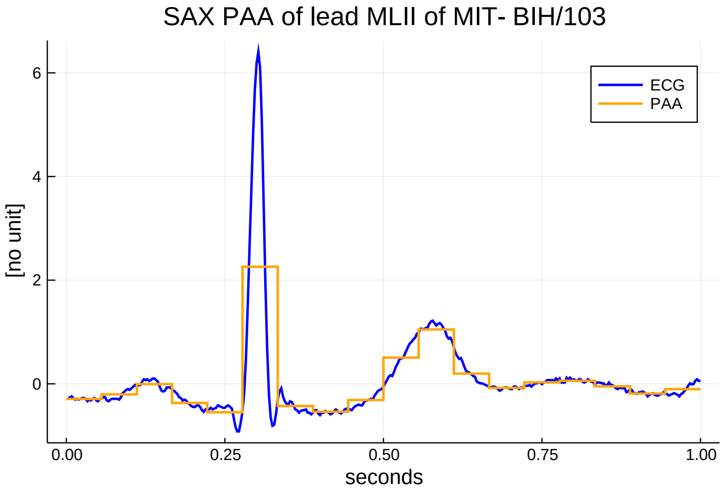


Figure 1: ECG with PAA, $w = 18$

SAX – Step (3)

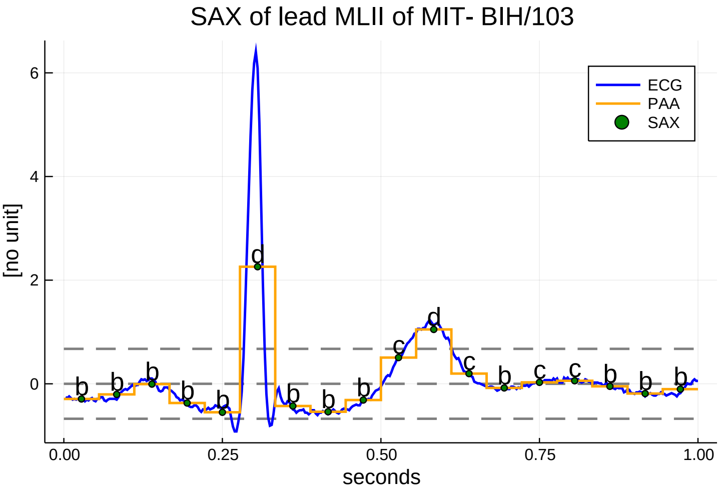


Figure 2: ECG with SAX, $w = 18$, $a = 4$

SAX and MSAX – Distance Measure

- needed to compare two SAX/MSAX segments
- sum of distances between symbols
- symbol distance is based on difference of breakpoints
- lower-bounds Euclidean Distance, i.e. corresponds to “real” distance

HOT SAX and HOT MSAX – Overview

- HOT SAX: find discords in SAX-represented time series
- classifies time series segments into “discord” and “non-discord”
- HOT MSAX: uses MSAX instead of SAX
- HOT MSAX can work with multivariate time series
- simple heuristic
- only two parameters

Analysis Process

(1) perform HOT SAX and HOT MSAX for many parameter combinations for all ECGs in the MIT-BIH database

(2) find recall, precision for each combination

(3) set recall threshold of 95%, then sort by precision

(4) choose top 10 of those parameters for each method

(5) choose best parameters for each method using box plot, interquartile range, outliers

Three Datasets

(1) S-SAX: data for HOT SAX algorithm, considering each lead separately

(2) D-SAX: data for HOT SAX, considering both leads combined

(3) MSAX: data for the HOT MSAX algorithm

Overview of Results

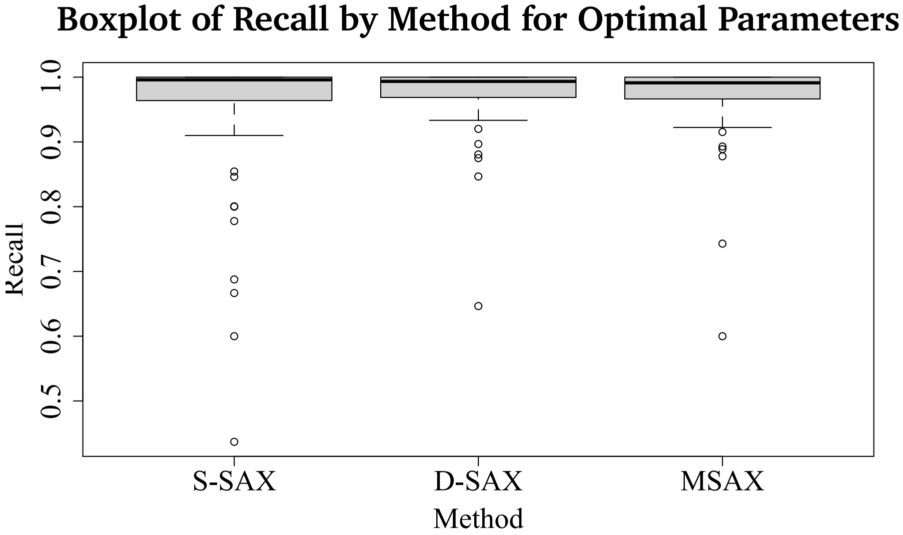
Table 1: Coarse Overview of Results. Shown are sets of parameters for each method that satisfy the recall threshold of 95%. ▸ More

Method	Total Sets	Sets Satisfying recall ≥95%
S-SAX	4,968	99 (1.2%)
D-SAX		192 (3.9%)
MSAX		255 (5.1%)

Best Parameter Sets by Method

Table 2: Best Parameter sets for each of the methods. Best overall parameters highlighted in bold.

Parameter Method	k	w, m	a
S-SAX	-1	36	21
D-SAX	-1	**12**	24
MSAX	-1	**12**	**17**


Boxplot of Recall by Method for Optimal Parameters
Recall
1.0
0.9
0.8
0.7
0.6
0.5
S-SAX
D-SAX
MSAX
Method

Comparing Best Parameter Sets – Recall

Table 3: Statistical measures for recall of optimal parameter sets. Best overall values highlighted in bold. ▸ More

Measure Method	IQR	Median	Outliers
S-SAX	0.035	**99.60%**	11
D-SAX	**0.030**	99.35%	**6**
MSAX	0.033	99.13%	**6**

Discussion

- no significant difference in recall for the methods

→ hypothesis that HOT MSAX has higher recall than HOT SAX cannot be supported

- MSAX has smallest alphabet size and highest dimension reduction

→ MSAX may be more efficient in achieving same results as S-SAX, D-SAX

- Anacleto *et al.* showed similar D-SAX vs MSAX performance which supports this work's conclusion

Novel Contributions

- application of MSAX to ECG discord discovery and medical data in general
- the HOT MSAX algorithm, a modification of HOT SAX that uses MSAX
- the expansion of HOT SAX to multivariate time series through HOT MSAX

Conclusion

- best parameters for S-SAX, D-SAX, MSAX were found
- could not demonstrate higher recall of HOT MSAX for best parameters
- showed viability of a discord classifier to assist with ECG analysis
- future research possibilities: use different ECG data, different detection criteria

References I

[1] *The top 10 causes of death*, 2020. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death` (visited on 05/25/2021).

[2] 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`.

[3] 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`.

[4] 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. DOI: `10.3934/mbe.2019105`.

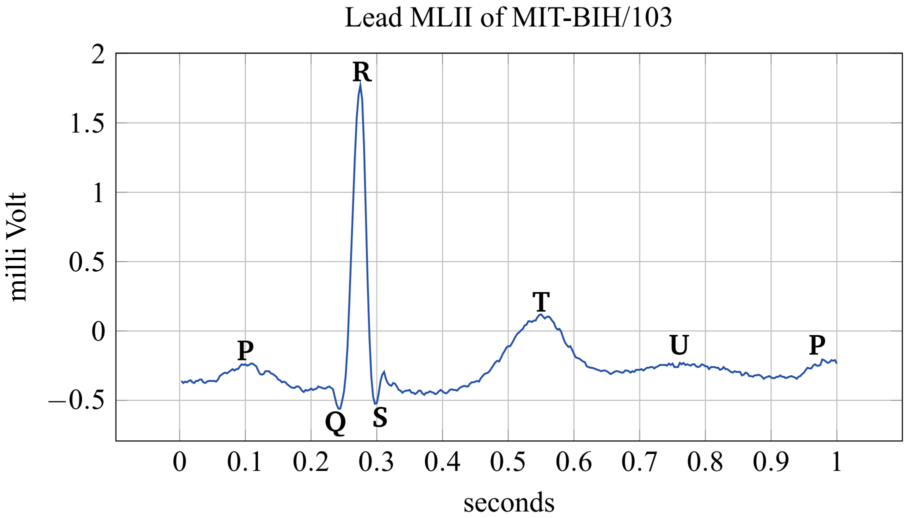
References II

[5] 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.

Thank You!

Appendix

Annotated ECG Graph



Background on Methods

- Lin *et al.* (2003):
  Symbolic Aggregate Approximation (SAX): simplified, symbolic representation

- Keogh *et al.* (2005):
  Heuristically Ordered Time series using SAX (HOT SAX): discord discovery algorithm using SAX

- Anacleto *et al.* (2020):
  Multivariate SAX (MSAX): expands SAX to multivariate time series

Time Series

Time series after [2]:

$$\{\mathbf{v}[t]\}_{t\in\{1,\ldots,T\}}$$

is a n-variate time series where for each time point t,

$$\mathbf{v}[t] = (v_1[t], \ldots, v_n[t])^T$$

are the values. T is length of time series, n is the number of values per moment

Univariate Normalization

Sample mean:

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{v}[t].$$

Sample standard deviation:

$$s = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{v}[t]-\bar{x})^2}$$

Normalization:

$$\mathbf{v}[t] = \frac{\mathbf{v}[t]-\bar{x}}{s}, \qquad \forall t \in \{1,\ldots,T\}.$$

Multivariate Normalization I

Mean vector:

$$E(\mathbf{V}[t]) = \vec{E} = \begin{bmatrix} \operatorname{mean}(V_1[t]) \\ \vdots \\ \operatorname{mean}(V_n[t]) \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} V_1[t] \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} V_n[t] \end{bmatrix}.$$

Covariance matrix:

$$\operatorname{Var}(\mathbf{V}[t]) = \Sigma_{n\times n} = \begin{bmatrix} \operatorname{cov}(V_1, V_1) & \dots & \operatorname{cov}(V_1, V_n) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(V_n, V_1) & \dots & \operatorname{cov}(V_n, V_n) \end{bmatrix}.$$

Multivariate Normalization II

Covariance Function:

$$\mathrm{cov}(V_i[t], V_j[t]) = E\Big(\big[V_i[t] - E(V_i[t])\big] \cdot \big[V_j[t] - E(V_j[t])\big] \Big)$$

Normalization following [2]

$$\mathbf{V}[t] = (\Sigma_{n\times n})^{-1/2} \left(\mathbf{V}[t] - \vec{E} \right).$$

Piecewise Aggregate Approximation

Following [3], PAA is calculated as

$$\overline{\mathbf{v}}[t] = \frac{w}{T} \sum_{j=\frac{n}{w}(t-1)+1}^{\frac{n}{w}t} \mathbf{v}[t], \qquad \forall t \in \{1, \ldots, w\}.$$

Performs smoothing and dimension reduction.

SAX Discretization I

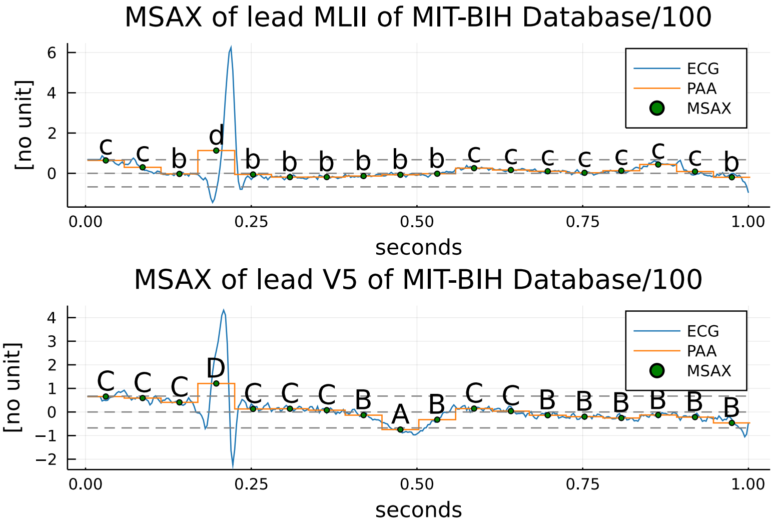
Idea: create a equiprobable symbols for discretization.
Action: split $\mathcal{N}(0,1)$ into a segments, use the resulting breakpoints.
Breakpoints are $B = \beta_1, \ldots, \beta_{a-1}$. The area under the normal curve $\mathcal{N}(0,1)$ (i.e. the probability) between two consecutive segments β_i and $\beta_{i+1} = 1/a$.

SAX Discretization II

Table 4: Breakpoint values for numbers of breakpoints a from 3 to 6. Table contents are quoted from [3].

a β_i	3	4	5	6
β_1	-0.43	-0.67	-0.84	-0.97
β_2	0.43	0	-0.25	-0.43
β_3	—	0.67	0.25	0
β_4	—	—	0.84	0.43
β_5	—	—	—	0.97

SAX Discretization III



MSAX of lead MLII of MIT-BIH Database/100

MSAX of lead V5 of MIT-BIH Database/100

SAX/MSAX Distance Measure I

SAX distance [3]:

$$MINDIST\left(\widehat{\mathbf{u}}[t],\widehat{\mathbf{v}}[t]\right) \equiv \sqrt{\frac{T}{w}}\sqrt{\sum_{t=1}^{w}\left(dist\left(\widehat{\mathbf{u}}[t],\widehat{\mathbf{v}}[t]\right)\right)^{2}}.$$

MSAX distance [2]:

$$MINDIST\_MSAX\left(\widehat{\mathbf{U}}[t],\widehat{\mathbf{V}}[t]\right) \equiv \sqrt{\frac{T}{w}}\sqrt{\sum_{t=1}^{w}\left(\sum_{i=1}^{n}\left(dist\left(\widehat{U}_i[t],\widehat{V}_i[t]\right)\right)^{2}\right)}.$$

SAX/MSAX Distance Measure II

dist function [3], is difference between breakpoints:

$$\text{cell}_{r,c} = \begin{cases} 0, & \text{if } |r-c| \leq 1 \\ \beta_{\max(r,c)-1} - \beta_{\min(r,c)}, & \text{otherwise} \end{cases}$$

SAX/MSAX Distance Measure III

Table 5: A table for the *dist* function for $a = 5$. Each cell displays the distance between the symbols denoting its row and column.

	a	b	c	d	e
a	0	0	0.59	1.09	1.68
b	0	0	0	0.51	1.09
c	0.59	0	0	0	0.59
d	1.09	0.51	0	0	0
e	1.68	1.09	0.59	0	0

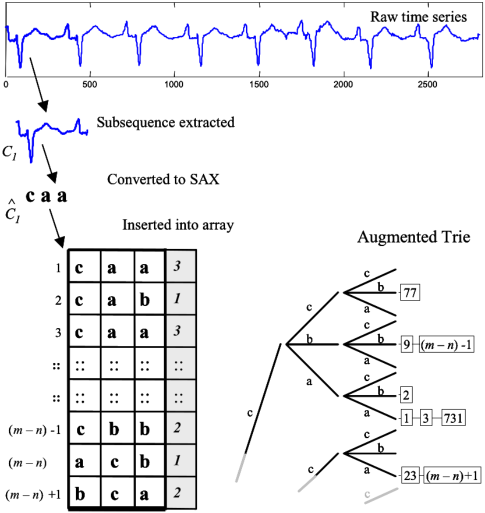
SAX/MSAX Distance Measure IV

Lower-bounding: infimum from set theory: largest value in set S smaller than all elements of a set $V \in S$.
For SAX, MSAX: MINDIST is smaller than Euclidean Distance (“true” distance), thus it is representative

HOT SAX/HOT MSAX Heuristic I

- two parameters: m and k
- two assumptions:
  - time series discords are rare
  - segments similar to discords may also be discords
- speed up discord discovery:
  - consider rarest segments first
  - consider similar segments together

HOT SAX/HOT MSAX Heuristic II



1	c	a	a	*3*
2	c	a	b	*1*
3	c	a	a	*3*
::	::	::	::	::
::	::	::	::	::
$(m-n)$ -1	c	b	b	*2*
$(m-n)$	a	c	b	*1*
$(m-n)$ +1	b	c	a	*2*

Statistical Analysis I

Table 6: Contingency table showing the relationship between detected discords and actual annotated values.

Assigned Actual	Discord Detected	Non-Discord Detected
Is Discord	True Positive	False Negative
Is Non-Discord	False Positive	True Negative

Statistical Analysis II

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}},$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}},$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}.$$

Finding Best Parameters – Dataset 1

Dataset 1 failed, thus dataset 2 was made

Table 7: Table of the methods used for dataset 1, the parameters of each method, the rationale behind the parameter choice, and the values the parameter takes are shown.

Method	Parameter	Rationale	Values
SAX/MSAX	w	arbitrary factors of 360	2,3,4,5,12,20,30,40,60
	a	arbitrary, $2\leq a \leq 25$	4,5,6,7,8,9,10,12,14,17,20
HOT SAX/MSAX	k	arbitrary	-1,25,50,100, 150,200,300,500
	m	arbitrary factors of 360 and of w	2,3,4,5,12,20,30,40,60

Finding Best Parameters – Dataset 1

Table 8: Results of the analysis of dataset 1. The total number of parameter sets and the number and proportion of parameter sets in dataset 1 that fulfill the analysis conditions are presented for each method.

Method	Total Sets	Sets Satisfying Analysis Conditions	
		recall $\geq$95%	recall $\geq$95% and $m \neq w$
S-SAX	2,640	3 (0.1%)	0 (0%)
D-SAX		13 (0.5%)	0 (0%)
MSAX		23 (0.9%)	3 (0.1%)

Finding Best Parameters – Dataset 2

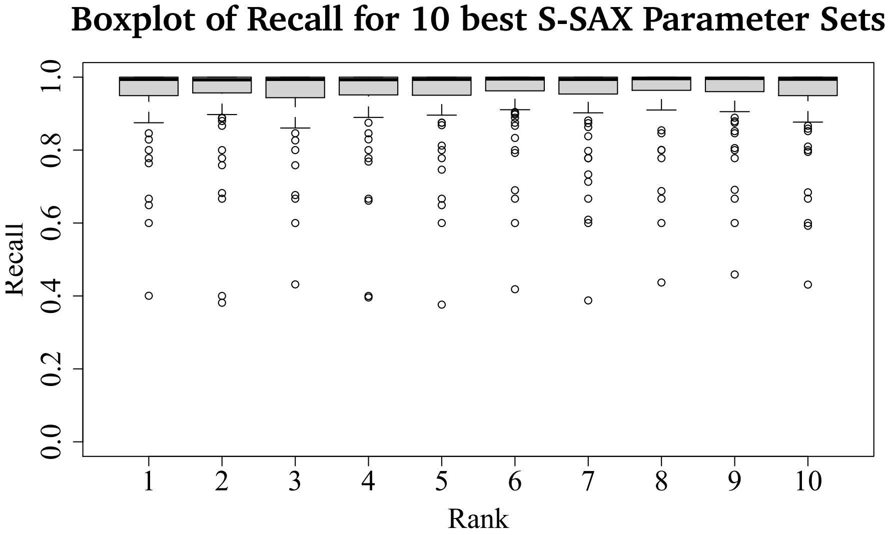
Table 9: Table of the methods used for dataset 2.

Method	Parameter	Rationale	Values
SAX/MSAX	w	factors of 360	2,3,4,5,6,8,9,10,12,15, 18,20,24,30,36,40,45, 60,72,90,120,180,360
	a	$2 \leq a \leq 25$, length of alphabet	$\overline{2,\ldots,25}$
HOT SAX/MSAX	k	arbitrary	-1,25,50,75,100, 150,175,200,300
	m	same as w	2,3,4,5,6,8,9,10,12,15, 18,20,24,30,36,40,45, 60,72,90,120,180,360

Finding Best Parameters – S-SAX

Table 10: Ranking of top 10 most precise S-SAX parameter combinations.

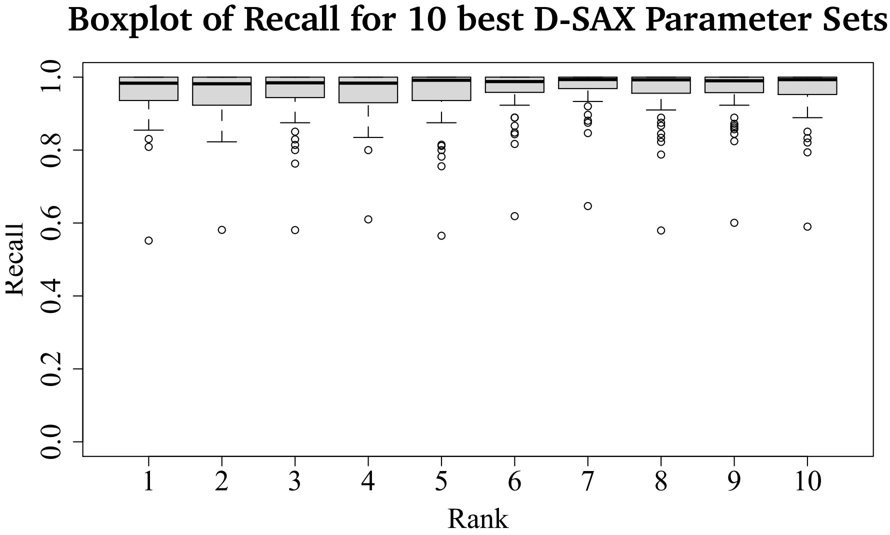
Properties Rank	k	w	m	a	Recall (%)	Accuracy (%)	Precision (%)
1	-1	40	40	19	95.21	37.46	35.94
2	-1	24	24	24	95.19	37.57	35.93
3	-1	30	30	22	95.37	37.46	35.93
4	-1	36	36	20	95.09	37.41	35.92
5	-1	45	45	18	95.24	37.33	35.89
6	-1	24	24	25	95.74	37.32	35.88
7	-1	72	72	15	95.02	36.86	35.87
8	-1	36	36	21	95.92	37.06	35.86
9	-1	30	30	23	95.72	37.17	35.86
10	-1	120	120	13	95.13	36.51	35.85


Boxplot of Recall for 10 best S-SAX Parameter Sets
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Rank
1
2
3
4
5
6
7
8
9
10

Finding Best Parameters – D-SAX

Table 11: Ranking of top 10 most precise D-SAX parameter combinations.

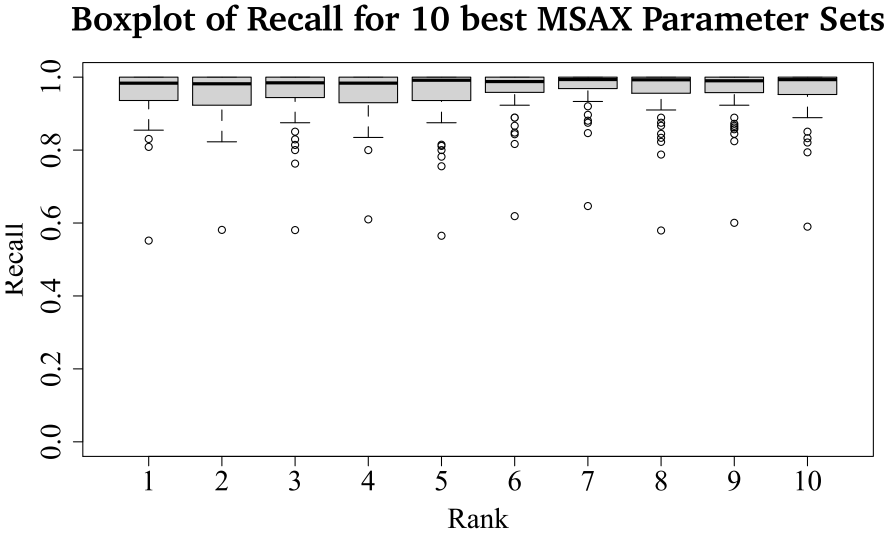
Properties Rank	k	w	m	a	Recall (%)	Accuracy (%)	Precision (%)
1	-1	12	12	22	95.28	41.91	36.56
2	-1	10	10	25	95.18	41.99	36.53
3	-1	15	15	19	95.14	41.70	36.46
4	-1	12	12	23	95.18	41.00	36.34
5	-1	40	40	12	95.08	39.64	36.29
6	-1	15	15	20	96.11	40.31	36.27
7	-1	12	12	24	97.04	40.16	36.26
8	-1	36	36	13	95.80	38.93	36.20
9	-1	20	20	17	95.87	39.82	36.19
10	-1	30	30	14	96.09	39.32	36.18


Boxplot of Recall for 10 best D-SAX Parameter Sets
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Rank
1
2
3
4
5
6
7
8
9
10

Finding Best Parameters – MSAX

Table 12: Ranking of top 10 most precise MSAX parameter combinations.

Properties Rank	k	w	m	a	Recall (%)	Accuracy (%)	Precision (%)
1	-1	6	6	24	95.37	40.68	36.24
2	-1	12	12	16	95.10	39.85	36.24
3	-1	9	9	19	95.20	39.70	36.13
4	-1	10	10	18	95.89	39.45	36.12
5	-1	8	8	21	96.01	39.53	36.12
6	-1	6	6	25	96.02	39.94	36.12
7	-1	36	36	10	95.16	38.47	36.08
8	-1	12	12	17	96.51	38.89	36.06
9	-1	30	30	11	95.49	38.26	36.03
10	-1	72	72	8	95.70	37.74	36.03


Boxplot of Recall for 10 best MSAX Parameter Sets
Recall
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
Rank

Finding Best Parameters – Comparison

Biserial Correlation Analysis was performed with the goal of identifying the influence of the methods. It was found that:
S-SAX vs D-SAX: 0.06
S-SAX vs MSAX: 0.033
D-SAX vs MSAX: -0.04