Applied Mathematics and Informatics Program

Multivariate Symbolic Aggregate Approximation for ECG Analysis

Moritz M. Konarski

A Thesis Submitted to the Applied Mathematics and Informatics Program of American University of Central Asia in Partial Fulfillment of the Requirements for the Degree of **Bachelor of Arts**

Author
Moritz M. Konarski

Certified by Thesis Supervisor
Professor Taalaibek M. Imanaliev

Accepted by
Sergey N. Sklyar
Head of Applied Mathematics and Informatics Program, AUCA

May 17, 2021

TODO Don't forget to put simply May, 2021 or something Bishkek, Kyrgyz Republic

ABSTRACT

Electrocardiograms are the most common tool used to diagnose heart diseases, which claim more lives each year than any other disease. Since their invention, CITE cite when, and who electrocardiograms needed to be analyzed by a trained professional like a cardiologist. Since CITE cite when it became a thing, they can be analyzed using computers and computer-assisted methods. These methods can be more accurate, faster, and more versatile than humans. One discord discovery method is the Symbolic Aggregate Approximation, which transforms an electrocardiogram into a shorter, symbolic form. This form is faster and simpler to analyze.

Multivariate Symbolic Aggregate Approximation takes more than one electrocardiogram lead into account and should thus be more accurate when it comes to discord discovery using Heuristically Ordered Time series using Symbolic Aggregate Approximation.

This paper shows, with TODO insert significance level, that Multivariate Symbolic Aggregate Approximation increases the sensitivity of HOTSAX compared to Symbolic Aggregate Approximation.

Keywords: acute cardiac ischemia, ECG, mathematical modeling

ACKNOWLEDGEMENTS

This research project is part of a Faculty Research Grant **TODO** insert specifications or ID here. I also want to thank my Supervisor, Professor Taalaibek M. Imanaliev, for his invaluable help and precise feedback.

TABLE OF CONTENTS

TODO Sections (not explicit):
- motivation
- objectives and contribution
- falsifiable hypothesis
- introduce relevance of sensitivity here

TODO Just start from scratch:
- introduction with who data for ihd, arrh
- prevention and detection is important
- ECG is THE method for that
- how an ECG works; pros and cons of an ECG
- how an ECG is also a multivariate time series
- where does automated ECG analysis come into play?
- time series methods to ECG
- discords that then lead to next point: connect to the 5 steps, signal which I will look at, where is SAX, HOTSAX, where is the cardiologist...

- the approach of sifting out the abnormal beats and having a cardiologist look at them instead of doing everything internally
- SAX, MSAX, HOTSAX and my hypothesis
- use definitions etc

1 INTRODUCTION

TODO UPDATE: In the year 2016, over 9.4 million people worldwide died of ischemic heart disease (IHD). IHD is responsible for 16.6% of all deaths, making it the most common cause of death globally. All forms of cardiovascular disease make up 31.4% of all deaths (17.9 million). Death caused by IHD disproportionally affects people over 50 years of age, with 91% of deaths for men and 95% of deaths for women occurring in that age range. In Kyrgyzstan, 13% of all deaths in 2016 were caused by IHD [1]. who2018

TODO FOCUS MORE ON MY ACTUAL TOPIC:

Ischemic heart disease is characterized by restricted blood flow to an area of the heart, causing it to not receive enough blood and oxygen. Blood flow restriction is caused by a blockage (or narrowing) in a blood vessel supplying the heart muscle. An artery can be blocked by a blood clot, but the most common cause is plaque buildup, which is called atherosclerosis. If the circulation to the heart is completely blocked, the cells in the heart muscle begin to die. This is called myocardial infarction, more commonly known as a heart attack. The deprivation of oxygen the heart experiences leads to the characteristic chest pain commonly associated with heart attacks [2]. iom2010

TODO mention arrhythmia too, it is what the mit database looks at

IHD can be diagnosed before it leads to a heart attack. The diagnosis can be performed based on a patient's medical history, pharmacologically induced stress, or stress induced by physical exercise. During an exercise stress test, an electrocardiograph (sometimes combined with other methods) records the patient's heart activity, resulting in an electrocardiogram (ECG) [2]. iom2010

TODO make this its own paragraph section

TODO

- heart's electrical activity
- up to 12 leads
- common medical diagnostic tool
- electricity is what causes the contraction
- this can be measured on the skin
- a bit on ECG theory
- specific electrodes and positions
- mention ion flow

The ECG is a diagnostic tool used to evaluate patients with (suspected) heart problems. It is a non-invasive, real-time, and cost-effective method that may be used to diagnose IHD. It is the most common tool used for cardiac analysis and diagnosis [3, 4, 5]. The most common form of the ECG is the 12-lead variant. The 12-lead ECG consists of 6 leads connected to the limbs and 6 leads connected to the torso of the patient. The leads record the differences in electrical potential between the places on the body that they are attached to. This reflects the differences in voltage that the heart experiences with each heart beat because those voltage differences are conducted by the body. The measurements are taken in millivolts (mV). The ECG represents the state of the heart; a recorded ECG has the shape of a wave (the ECG wave) [4, 5].

TODO datasets are available online, the most significant leads tend to be included

If the state of the heart beat changes as the result of a disease like IHD (changing the measurable potentials or their occurrence over time), the ECG is able to record these changes.

The characteristic shape of an ECG for two heart beats is shown in Figure 1.1; the figure is taken from [6]. The figure has been annotated to show the significant features of an ECG. The peaks (or waves) P, Q, R, S, T, and U, as well as the segments between them, are the focus of ECG analysis. Multiple points together form what is called a complex; TODO FIX: the QRS complex is a good example of this. Using these waves, the heart activity can be described and analyzed. In an ECG, the P-wave is the result of the atria depolarizing, which is the process of blood entering the heart as the first step in a heart beat. The QRS complex represents ventricular depolarization, the contraction of the heart causing it to pump blood. The T-wave is the return of the ventricle to its polarized state. The U-wave is only present in roughly 25% of the population and may be caused by mechanical-electric feedback. The RR interval can be used to calculate the heart rate because it represents one complete heart beat [6]. The shape of the P, Q, R, S, T, and U waves as well as the duration of various intervals between them are used as indicators of cardiac diseases.

TODO UPDATE TO TIKZ FIGURE:

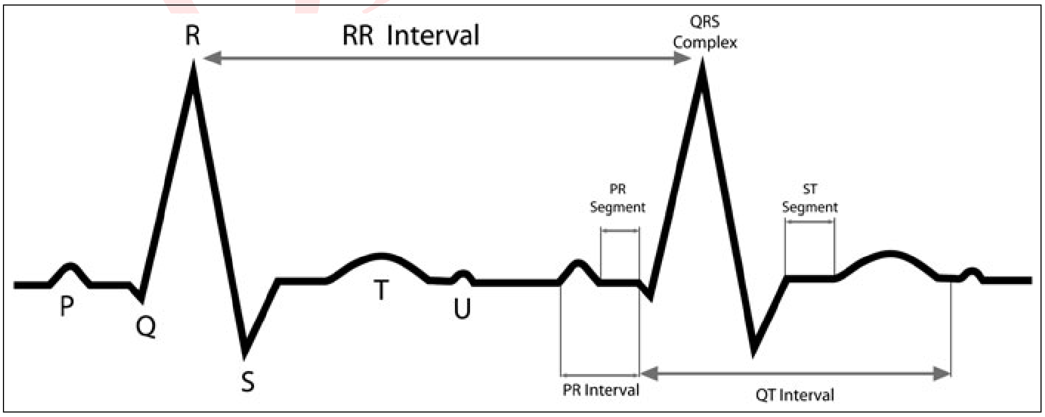

Figure 1.1: A schematic of an ECG waveform, annotated; from [6]

TODO also: lots of data, cannot be analyzed quickly

Using an ECG to diagnose a cardiac condition is difficult in practice. Small changes in the components of the ECG can be indicators of diseases and those changes can be overlooked, even by trained and specialized physicians. The chance to make a mistake is even higher for non-specialized physicians and trainees [3, 5].

TODO make it about assisting in their diagnosis by pointing out important segments -> discords

For the diagnosis of IHD, changes in the ST-segment and T-wave are of particular interest. An elevation of the ST-segment compared to a normal heart beat is one of the main indications of IHD and myocardial infarction. A downward depression of the ST-segment, especially in combination with chest pain, is another indication of IHD. The changes in the ST-segment are thought to be caused by current flow between healthy heart muscle and ischemic heart muscle [6, 7].

TODO after introducing idea of discord discovery, introduce ECGs as time series:
- ECGs are multivariate: 2-12 leads
- discrete: measured at discrete points
- ordered sequences: come one after another after equal time segments
-

TODO then, time series analysis methods are ready to be applied to ECGs

TODO what is required of the method:

- fast
- accurate
- adaptable
- versatile

The diagnosis of IHD on the basis of an ECG is time sensitive. If a patient has IHD or suffers from a heart attack, treatment has to be started as soon as possible. Some forms of treatment are most effective in the first 3 hours after symptom onset and lose most of their effectiveness after 9 to 12 hours. The diagnosis required for treatment to begin should thus be as quick as possible. The ECG delivering information in real-time is an advantage here, even though there are more time consuming methods that can deliver more accurate results than an ECG [8].

TODO WHY automated ECG analysis:

The widespread use of ECGs and the time-sensitive nature of their application as diagnostic tools makes errors, delays, or inconsistencies in their interpretation unacceptable. A recent approach to minimizing this problem is the application of computer technology in ECG recording, storage, and analysis. The main steps of computerized ECG analysis are [4] TODO fix this one, adapt to my writing (1) signal acquisition and filtering, (2) data transformation or preparation for processing, (3) waveform recognition, (4) feature extraction, and (5) classification or diagnosis.

This research will investigate steps (3) and (4) through the use of different feature extraction algorithms. The ECG data will be retrieved from the European ST-T Database. This database provides ECG recordings that can be used as trial data to test feature extraction algorithms. The European ST-T Database contains annotations made by cardiologists indicating the ST-segment, T-wave, and their changes. They also include information about the suspected disease [9, 10]. This information can be used to determine the effectiveness of the feature extraction algorithms.

TODO Sections:

- overview of methods for ECG analysis
- main elements
- current foci
- use the confusion matrix at all?
- support assertions made in introduction
- arrive at natural conclusion that SAX/MSAX/HOTSAX should be investigated
- describe all 4 statistical measures in some detail

TODO Structure:

- follow the same overall structure as the introduction, order of topics
- background on history of ECG
- how an ECG works in more technical terms
- how do cardiologists detect heart diseases?
- move the annotated graph down here and leave the general graph in intro?
- which time series methods are being applied to ECGs? strengths an weaknesses
- what are current hot topics in this field?
- more research on sax, msax, hotsax and what people have done with it
- ecg applications; why is it relevant
- RESEARCH MORE ABOUT ”ecg discord discovery algoorithm”
- use some of their arguments to support my choice
- try to make it flow so that my choices seem to come from the literature analysis
- maybe add a section that talks about how to evaluate these types of algorithms

2 STATE OF COMPUTERIZED ECG ANALYSIS

Recent advances in computer technology have enabled the use of computers in every aspect of ECG acquisition, processing, analysis, and storage. In light of these developments, the American Heart Association published recommendations for the interpretation and standardization of the ECG. They recommend that the low-frequency cutoff for low-frequency filtering of an ECG should be 0.05 Hz or 0.67 Hz for filters that do not exhibit phase distortion. For high-frequency filtering they recommend a cutoff of at least 150 Hz. For the storage of digital ECG samples (at 500 samples per second), it is recommended use use compression with an error of less than 10 microvolt [4].

Xie *et al.* [5] provide an overview of the current approaches to computerized ECG analysis. The standard approach to using computerized methods in ECG analysis is comprised of four steps (1) denoising of the raw ECG signal(s), (2) feature engineering, (3) dimensionality reduction, and (4) classification. To denoise an ECG, digital filters are often used. Their drawbacks are that they only filter out very specific frequencies. Because noisy ECGs contain different types of contaminations, digital filters can be inaccurate. Using wavelet transforms for denoising has the advantage that noise can be more precisely targeted and the clean signal reconstructed afterwards. Choosing appropriate wavelet parameters can be challenging and methods to optimize this process have been proposed. Empirical mode decomposition is the third option generally employed to denoise an ECG. It does not require the user to set parameters but it can lead to a mixing of oscillations of different time scales.

After the signal has been appropriately denoised, feature engineering is performed. Feature engineering is the process of extracting features that are relevant for diagnosis from the many points the ECG signal contains. The main features targeted for extraction are the PQRST features mentioned in the introduction. The fast Fourier Transform provides a way of analysing the frequency domain of the ECG signal, enabling the detection of the QRS complex and other features. The missing time information in the fast Fourier Transform can lead to difficulties in detecting time-dependent features. The short-time Fourier Transform adds time information to the fast Fourier Transforms data. This can increase the accuracy of the feature extraction. This transform has the drawback that there is a tradeoff between the time and frequency resolutions. Wavelet transforms can also be used for feature extraction. They have the advantage that they are suitable for all frequency ranges. Choosing the right wavelet base for the desired application can be a challenge. The discrete wavelet transform is the most widely used wavelet transform, thanks to its computational efficiency. Statistical methods are also used to extract features from ECGs; those methods are generally less affected by noise in the signal.

After the features of the ECG have been extracted, it is often necessary to reduce the number of features. The reason for this is that a large number of features, despite their high accuracy, require a high amount of computation to classify. This lengthy computation can negate the advantages gained by high accuracy. This process sacrifices a certain amount of information and sometimes precision, but significantly speeds up the classification. Feature selection is a process that attempts to select a subset of the original data that adequately describes the whole data. Feature selection can be performed by a filter that filters out unnecessary attributes based on some metric. This methods is relatively simple, but the filtering process removes data and thus negatively impacts the precision of further steps. Feature extraction on the other hand uses dimensionality reduction methods to keep as much of the original information as possible. Principal component analysis preserves as much of the variance in the original data as it can. Other algorithms focus on separating classes of data, pattern recognition, or retaining the structure of the original data.

The final stage of the ECG processing is the classification stage. In this stage judgements are made based on the prepared input data and the result should be a disease diagnosis. In the early stages of computerized ECG analysis classification was performed by algorithms based on human actions when reading an ECG. Those algorithms were basic and not particularly accurate. Currently, the classification at the end of the preparation process is performed by a machine learning algorithm. Such models include the k-nearest-neighbors model which classifies points into groups but which is very expensive to calculate for high-dimensional data. Support vector machines are used for pattern recognition and are able to work with small samples. Artificial neural networks are robust and can work with complex problems, they are generally more accurate than support vector machines. The newest approach is to forego the stages discussed here and use a single neural network to perform all the required tasks ”end-to-end”. These networks are fed raw data and the denoising, feature extraction, selection, and classification is performed internally by the model [5].

The end-to-end approach to ECG analysis is a relatively new development and is being actively researched. The more traditional method using denoising, features engineering, and classification as separate steps is also still relevant. The combination of denoising and feature extraction with a machine learning classifier can lead to very good results. Prasad and Parthasarathy [11] use the fast Fourier Transform to extract features from an ECG and then employ a multi-objective genetic algorithm to detect abnormal ECG signals with high accuracy. Vaneghi *et al.* [12] compare 6 common feature extraction techniques with respect to their detection of ventricular late potentials. The compared methods are the autoregressive method, wavelet transform, eigenvector, fast Fourier Transform, linear prediction, and independent component analysis. Valupadasu and Chunduri [13] use the fast Fourier Transform to analyze the energy level in different frequencies in the ECG of patients with IHD. They find that the energy is distributed differently, allowing the distinction of ECGs with IHD from those without IHD. Kaur, Rajni, and Marwaha [14] analyzed ECG signals with both the wavelet transform and principal component analysis. They found that the wavelet transform outperformed principal component analysis for the detection of heart beats in an ECG. Their model achieved an error rate of 0.221%

of incorrectly classified heart beats .

3 METHODS

TODO Sections:
- general overview -> flowchart
- then explain each element of the flowchart one by one
- use formulae etc
- nice amount of tikz graphs
- section on implementation with details and the more important elements use another flow chart?
- use graphs to illustrate all important elements
- make a data description section that describes my process of data handling; which database

- explain the parameters that the methods have and what they mean
- describe how a got all the data

This section explains the methods used in this research. TODO create flow charts for all this shit to make it simpler. First methods section for the analytical methods in a mathematical way.

3.1 Mathematical Foundations

- section for the workings
- explain theoretical foundations of the approach
- what is it grounded in, who, what, when

3.2 SAX

- idea
- normalization
- dimensionality reduction
- discretization
- distance measure
- TODO all with graphs and formulas

3.3 MSAX

- idea
- normalization
- dimensionality reduction
- discretization
- distance measure
- TODO all with graphs and formulas
- TODO points out differences to SAX

3.4 HOTSAX

- what is hotsax
- its theoretical foundations
- advantages, disadvantages
- how does it work

3.5 Statistical Analysis of Results

- explain true positive, true negative, and so on
- explain recall, accuracy, precision, f1
- explain why recall was chosen and if that is fair
- introduce the correlations that we would expect to find if my hypothesis is true and also the ones that would disprove it
- which types of correlation, significance testing, and modeling will be used and why; what are the justifications

3.6 Implementation

How I implemented the above stuff. Languages, approaches, hurdles, all the details needed to reproduce this research. Also mention the simplifications I chose to make and why: no sliding window, only even divisors, only divisors within sampling frequency and cutting ECG to even multiple of sampling frequency.

3.6.1 ECG acquisition

flow chart for process

- where to download
- what exactly are the ECGs
- where do they come from
- technical parameters of them
- the physionet suite
- annotations, what they mean, how I can get them, etc

3.6.2 preprocessing

flow chart for process

TODO the codes and constants given for each thing

- how were they preprocessed
- physionet suite
- my script and what it does and why
- problems and limitations of this
- libraries used

3.6.3 SAX

TODO how was the whole data thing handled, how is the data created

flow chart for process

- how was sax implemented
- how does HOTSAX work here
- libraries used

3.6.4 MSAX

flow chart for process

- how was sax implemented
- how does HOTSAX work here
- TODO point out differences to SAX
- libraries used

3.7 Statistical Evaluation

- reading the data into R
- summarizing the data
- the summarized data files
- libraries used

4 RESULTS

TODO Sections
- use confusion matrices for what is is vs what was predicted [p. 44 anacleto2019]
- compare all the parameters and their influence
-

4.1 First Run

- parameters for this run
- why could I not let it continue
- what did this run indicate -> what did I change and modify for the next run
- keep in mind that the lower recall can be caused by the way I do the ECG checking, and that I did not want to assign data to segments that did not have it before out of fear that I would invent results.

4.2 Second Run

4.3 SAX

influence and significance of all the major parameters:
- k
- paa count
- subsequence count
- alphabet size
- which ones seem to be the best

4.4 MSAX

influence and significance of all the major parameters:
- k
- paa count
- subsequence count
- alphabet size
- which ones seem to be the best

4.5 MSAX vs SAX

Comparing SAX to MSAX is done using the recall value defined in TODO reference. Investigating the correlation between the methods (represented by a 1 for SAX and a 0 for MSAX), yields the correlation coefficient of -0.25. This coefficient indicates that for all investigated parameter combinations, the use of the MSAX method is weakly correlated with an increase in recall. When a specific set of parameters is selected and the correlation analysis is repeated, the correlation coeficient is -0.73, indicating a strong correlation. Here $k = -1$ and paa_count = 12.

- just the results that are gained directly from the data
- put results in graphs and tables to make them referencable

5 DISCUSSION

5.1 MSAX vs SAX

TODO use the results from the previous section to come to a conclusion

- do proper hypothesis testing of my hypothesis statement
- argue which sets of parameters are the most effective
- judge if I proved what I set out to prove
- what are the uses of this method
- which applications could this fit?
- what should be done in future research

REFERENCES

who2018 [1] World Health Organization, *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016*. Geneva, 2018. [Online]. Available: https://www.who.int/healthinfo/global_burden_disease/estimates/en/.

iom2010 [2] Institute of Medicine (US) Committee on Social Security Cardiovascular Disability Criteria, *Cardiovascular Disability: Updating the Social Security Listings*. Washington (DC): National Academies Press (US), 2010, pp. 101–131. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK209964/.

alghatrif2012 [3] M. AlGhatrif and J. Lindsay, "A brief review: History to understand fundamentals of electrocardiography," *J Community Hosp Intern Med Perspect*, vol. 2, no. 1, 2012. DOI: 10.3402/jchimp.v2i1.14383.

kligfield2007 [4] P. Kligfield *et al.*, "Recommendations for the standardization and interpretation of the electrocardiogram: Part I: The electrocardiogram and its technology: A scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: Endorsed by the International Society for Computerized Electrocardiology," *Circulation*, vol. 115, no. 10, pp. 1306–1324, 2007. DOI: 10.1161/CIRCULATIONAHA.106.180200.

xie2020 [5] L. Xie *et al.*, "Computational Diagnostic Techniques for Electrocardiogram Signal Analysis," *Sensors (Basel)*, vol. 20, no. 21, 2020. DOI: 10.3390/s20216318.

asilewski2012 [6] J. Wasilewski and L. Poloński, "An introduction to ECG interpretation," in *ECG Signal Processing, Classification and Interpretation*, A. Gacek and W. Pedrycz, Eds. London: Springer, 2012, ch. 1, pp. 1–20. DOI: 10.1007/978-0-85729-868-3_1.

autaharju2009 [7] P. M. Rautaharju *et al.*, "AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part IV: The ST segment, T and U waves, and the QT interval: Ascientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: Endorsed by the International Society for Computerized Electrocardiology," *Circulation*, vol. 119, no. 10, e241–250, 2009. DOI: 10.1161/CIRCULATIONAHA.108.191096.

herring2006 [8] N. Herring and D. Paterson, "ECG diagnosis of acute ischaemia and infarction: Past, present and future," *QJM: An International Journal of Medicine*, vol. 99, no. 4, pp. 219–230, 2006. DOI: 10.1093/qjmed/hcl025.

physionet [9] A. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, 2000. DOI: 10.13026/C2NK5R.

taddei1992 [10] A. Taddei *et al.*, "The European ST-T database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography," *Eur Heart J*, vol. 13, no. 9, pp. 1164–1172, 1992. DOI: 10.1093/oxfordjournals.eurheartj.a060332.

prasad2018 [11] B. V. P. Prasad and V. Parthasarathy, "Detection and classification of cardiovascular abnormalities using FFT based multi-objective genetic algorithm," *Biotechnology & Biotechnological Equipment*, vol. 32, no. 1, pp. 183–193, 2018. DOI: 10.1080/13102818.2017.1389303.

vaneghi2012 [12] F. M. Vaneghi *et al.*, "A comparative approach to ECG feature extraction methods," in *2012 Third International Conference on Intelligent Systems Modelling and Simulation*, 2012, pp. 252–256. DOI: 10.1109/ISMS.2012.35.

alupadasu2012 [13] R. Valupadasu and B. R. R. Chunduri, "Identification of cardiac ischemia using spectral domain analysis of electrocardiogram," in *2012 UKSim 14th International Conference on Computer Modelling and Simulation*, 2012, pp. 92–96. DOI: 10.1109/UKSim.2012.22.

kaur2016 [14] I. Kaur, R. Rajni, and A. Marwaha, "ECG signal analysis and arrhythmia detection using wavelet transform," *Journal of The Institution of Engineers (India): Series B*, vol. 97, no. 4, pp. 499–507, 2016. DOI: 10.1007/s40031-016-0247-3.