1 BACKGROUND AND RELATED WORK

This section provides background information on time series and ECGs, as well as methods to analyze them.

1.1 Time Series and Time Series Analysis

This subsection will provide background information on time series and time series analysis methods. A time series is a set of values recorded at specific times. A common form of time series are discrete-time time series (often simply called discrete time series). Discrete time series are time series whose values are recorded at discrete points in time, the most common example of this are time series with values recorded at fixed intervals. Continuous-time time series are time series that are recorded continuously over a certain interval [1]. Time series that contain a single value for each moment in time are called univariate time series, while time series that record multiple values at each moment in time are called multivariate time series [2]. Time series are used in many disciplines to record information on time-dependent processes, e.g. stock prices in economics, the sun's activity in physics, or the heart's activity in medicine. Time series can be recorded digitally, physically, or, if they were recorded physically, can later be digitized. The recorded data can then be used to gain insight into the processes that were studied. To gain insight using a time series, the relevant information needs to be extracted from it–a process that is often called data mining. Data mining of time series is a vast discipline that, among others, includes [3, 4]:

- visualization (graphical representation),
- forecasting (predicting future behavior),
- indexing (finding the most similar time series to a given one),
- clustering (dividing time series into groups of similar ones),
- anomaly detection (detecting parts that are not "normal" or do not fit certain parameters),
- classification (assigning a label based on its features, e.g. "sick" and "not sick"), and
- summarization (reducing the complexity–often length–while preserving important features).

Challenges for time series analysis include the often very large data sets that are difficult for humans to analyze and take up considerable digital storage space. Analyzing very large data sets requires a large amount of computational power because most data mining algorithms become less efficient with larger data sets [3]. To mitigate this issue, time series dimension reduction (also known as dimensionality reduction or time series representation) is used. Dimension reduction transforms a "raw" (unmodified) time series into a representation that is simpler but nonetheless resembles the raw time series. This can be achieved by either using a method that reduces the number of values in a time series, or by extracting only the relevant features from the time series. According to [4, 5], there are four types of dimension reduction methods:

1. data dictated,
2. non-data adaptive,
3. model-based, and
4. data adaptive.

Methods 2–4 have their dimension reduction factors set by user-defined parameters. This means that the user can determine how much the dimension of the data should be reduced [4].

1.1.1 Data dictated representation

Data dictated methods derive their compression ratios from the data automatically, the most common form of this method is the clipped representation [4]. This representation simply transforms the raw time series into a sequence of 1s and 0s. A data points is assigned a 1 if its values is larger than the mean value of the time series, and a 0 otherwise. A sequence of 1s and 0s can be further compressed using various methods from computer science, finally yielding a very large compression ratio of 1057:1 [6].

1.1.2 Non-data adaptive representation

Non-data adaptive methods operate on time series segments with a fixed size to reduce the dimension and they are useful for comparing multiple time series with each other. These methods include the Discrete Wavelet Transform (DWT), the Discrete Fourier Transform (DFT), and the Piecewise Aggregate Approximation (PAA) [4]. The DWT uses wavelets, a limited-duration wave with an average value of 0, which represents both time and frequency information. The DWT is calculated using a series of filters applied to the signal. In [7], the DWT is used to detect beats in ECG signals and achieves a 0.221% detection error rate. The Fast Fourier Transform, an optimized form of the DFT, decomposes the its input signal into many sinus waves of different frequencies. In [8] it is used in conjunction with a machine learning model to achieve a beat classification accuracy of 98.7%. The PAA is part of the process of the SAX representation, thus it will be covered in **TODO** refer to the appropriate methods section

1.1.3 Model-based representation

Model based methods use stochastic methods such as Hidden Markov Models (HMM) and the Auto-Regressive Moving Average (ARMA) [4]. A HMM was used in [9] to cluster electroencephalograph recordings (measuring the brain's electrical activity). It was found that their methods was competitive with other established methods in classifying electroencephalograph signals. An auto-regressive model can be used to correctly identify a specific type of arrhythmia in an ECG and to group the occurrences of this arrhythmia together [10].

1.1.4 Data adaptive representation

Data adaptive methods use non-fixed size segments and aim to fit the raw data most closely. Examples of data adaptive methods are the Piecewise Polynomial Approximation (PPA), Piecewise Linear Approximation (PLA), Piecewise Constant Approximation (PCA), and SAX [4]. PPA can be used to compress and ECG by approximating it using polynomials. With second-order polynomials, ECGs can be compressed with a minimal level of distortion [11]. The authors of [12] use a modified PLA representation with adaptive ECG segmentation to successfully reconstruct the 12 standard leads of an ECG from only 3 leads. Using adaptive PCA as the dimension reduction method, the preprocessing and segmentation of ECGs can be significantly sped up while maintaining accuracy comparable to precious methods [13]. The SAX representation will be covered in detail in **TODO** refer to the SAX section and the following subsection 1.1.5 will provide background on the method and its variations.

1.1.5 SAX representation background

A particular dimension reduction method is SAX. Introduced by Lin, Keogh, Lonardi, and Chiu, SAX is a symbolic time series representation method for univariate time series. The authors felt that the symbolic methods available in 2003 did not provide the desired dimension reduction, did not correspond to the raw data accurately enough, and could not be applied to a subset of the total data. SAX uses the averaging of a user-defined number of segments and the labeling of segments with letters to reduce the dimension of the time series data. The number of letters, called the alphabet size, can also be chosen by the user and influences the dimension reduction. The distance between two time series in the SAX representation is guaranteed to resemble the distance between the two raw time series, this is called the distance measure. Since its creation, SAX has found widespread use in data mining and many researchers have attempted to modify and improve it.

The SAX distance measure has been improved to include the standard deviation [14] and a measure of the trend of each averaged segment [15, 16]. Extended SAX modifies SAX to include the minimum and maximum values of each segment for improved representation of the raw data [17] while 1d-SAX incorporates a linear regression over each segment into SAX [18]. A combination of SAX and a polynomial approximation was used to speed up the SAX method [19]. To improve the indexing performance of SAX, iSAX introduced convertible alphabet sizes, allowing SAX representations with different alphabet sizes to compared with each other and indexed into a tree structure [5]. iSAX 2.0 improves the iSAX index by reducing its computational complexity, enabling it to index a time series that has one billion elements, something that SAX or iSAX cannot do [20]. To perform time series anomaly detection using SAX, Keogh, Lin, and Fu introduced Heuristically Ordered Time series using SAX (HOT SAX) in 2005. Specifically, the authors attempt to detect time series “discords”, a subsequence of a time series that is most different from other segments of the time series. This can theoretically be done by simply comparing all subsequences of the raw time series to all other segments, but this approach is not feasible for long time series because of its complexity. Thus, HOT SAX utilizes SAX to reduce the dimensionality and complexity of the time series and then sorts the resulting SAX segments to speed up the discord detection. The authors suggest further research to investigate the use HOT SAX on multivariate time series [21]. For an in-depth description of this method, please refer to **TODO** refer to the methodology section of HOT SAX.

TODO mention this in the methodology section as supporting information

SAX and its variants have also been used for the analysis of multivariate time series. SAX-ARM combines the SAX representation with association rule mining (identifying rules and implications found in the data, i.e. parameter a influences parameter b) to analyze multivariate time series and discover the rules underlying the data [22]. Anacleto, Vinga, and Carvalho introduced MSAX in 2020 and thus expanded the use of SAX to multivariate time series. They utilize multivariate normalization with a covariance matrix and a modified distance measure to achieve this. To analyze their method, the authors use MSAX and SAX in a classification task based on multiple multivariate time series data sets. For these multivariate data sets, SAX was applied to each of their individual time series and those results were combined. Their analysis found that, overall, SAX applied in this way is superior to MSAX when it comes to classification accuracy. In 6 of the 14 tested data sets, SAX was significantly more accurate, in 2 of the MSAX was more accuarate, and in the remaining 6 their performance was not significantly different. It should be noted that in the ECG data set they tested, the accuracy of SAX (~87%) was slightly higher than that of MSAX (~84%), but not significantly so. Anacleto, Vinga, and Carvalho suggest that in future research MSAX should be applied to electronic health records (e.g. ECGs) and that it should be applied to other time series data mining applications besides classification [2]. MSAX will be thoroughly presented in **TODO** refer to methods section. **TODO** mention this in the methodology section as supporting information Another application of SAX to multivariate data used it to visualize multivariate medical test results and enable their analysis [23]. Resource-aware SAX is a SAX variant developed to analyze ECG using a mobile device like a mobile phone. The method takes advantage of the computational efficiency of SAX to perform the ECG analysis on the device and even preserve its battery life. Another application of the SAX method to ECGs is [24], which uses SAX with an added binary measure of the trend of each segment to detect ECG anomalies, achieving a recall value of 98%. The section 1.2 below will elaborate on ECGs and methods of their analysis.

1.2 ECGs and ECG Analysis

The following subsection covers the ECG and methods used in its analysis. Luigi Galvani noted the electrical activity in muscles 1786, but the history of the ECG only started in 1842, when Carlo Matteucci showed the electrical activity of a frog's heartbeat. In the 1870s, it was discovered that each heartbeat is characterized by electrical changes. Then, in 1901-1902, Willem Einthoven created the first ECG recording of a human heartbeat using using 3 leads connected to the limbs of the patient. Einthoven was the first to publish an ECG waveform with the now standard annotations P, Q, R, S, and T for the different features (see Figure 1.1). He would receive the 1924 Nobel Prize in medicine for his invention of the electrocardiograph. As a result of further development, the 12-lead ECG that we know today was created [25, 26].

TODO add a list of the leads and where they are

An ECG records the electrical activity that accompanies the contraction and relaxation of the heart muscle. The sinuatrial node, which can spontaneously give off an electrical pulse, initiates the heart beat. Its pulse is conducted through the heart by other specialized fibers, causing the heart to beat. The conduction of electricity is facilitated by Sodium, Calcium, and Potassium ions flowing in and out of cardiac cells [27]. Figure 1.1 shows a ECG wave of a single heartbeat from record 103 of the MIT-BIH data base [28, 29] (for more information on the data base, see section **TODO** refer to the ECG data base section). The P wave is caused by the depolarization of the atrial node, which

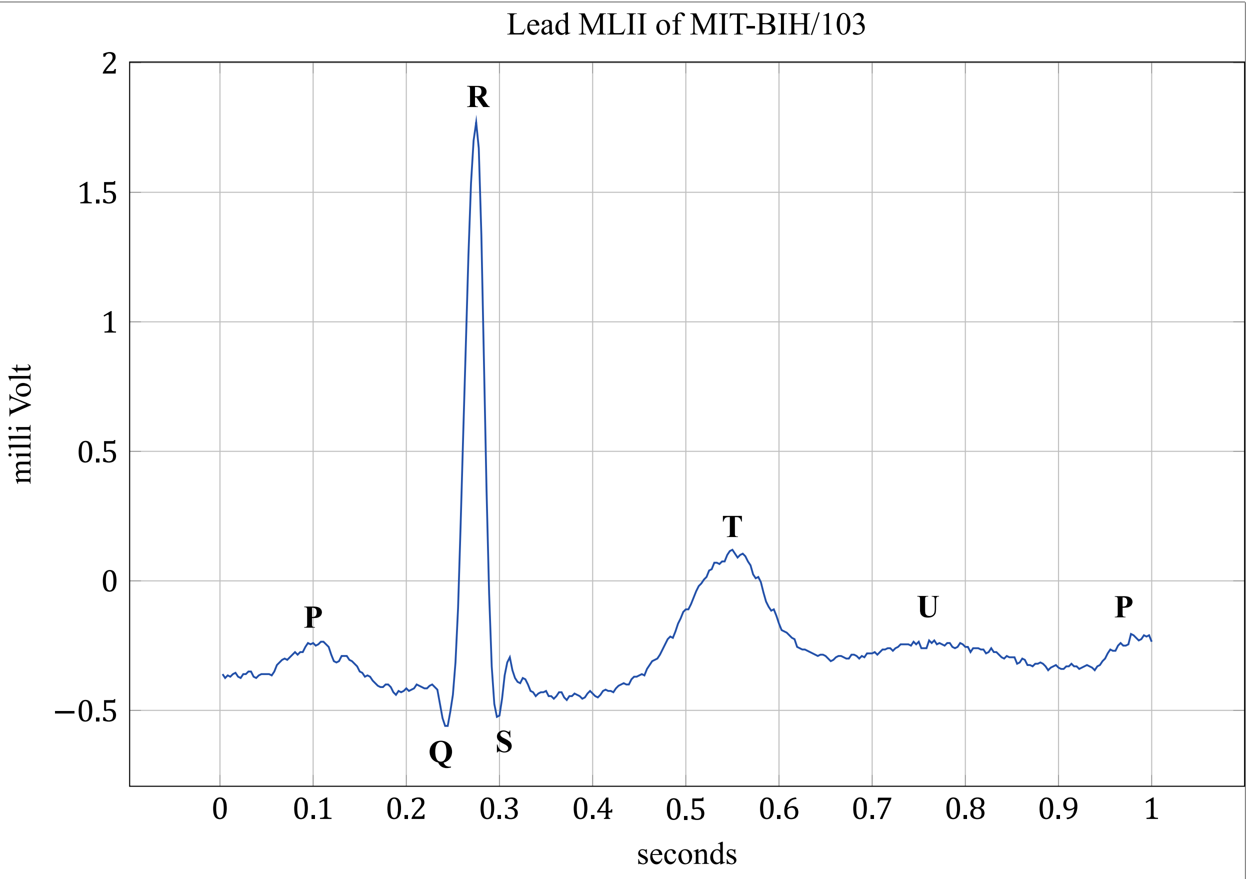

Figure 1.1: Annotated ECG of one heartbeat. This graph is based on lead II, data points 2031–2390 of recording 103 of the MIT-BIH data base [28, 29].

allows blood to flow into the heart. The QRS complex, as it is called, is the result of ventricular depolarization and represents the action of pumping blood out of the heart. The T wave is caused by ventricular repolarization in preparation for the next heartbeat. The U wave, only present in about 25% of people, is thought to he caused by mechanical-electric feedback [27, 30]. The last P wave is part of the next heartbeat, which is not shown in Figure 1.1.

The waves and complexes shown in Figure 1.1 are the object of ECG analysis. Changes in their shape, duration, or height can indicate heart conditions. Becker list some of the features relevant for ECG analysis [27]:

- The regularity of the rhythm: are the intervals between the QRS complexes and P waves regular?
- The shape of the QRS complex: do they have similar shape and duration?
- The regularity of the P waves: are the P waves similar and is the interval between P wave and QRS complex similar?
- Is the heart rate regular: measuring the time between QRS complexes can be used to calculate the heart rate, is this heart rate in the normal range?
- Do the waves and complexes come in the same order each time: each cycle should consist of a P wave, QRS complex, T wave.

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 [25, 31].

TODO inset some stuff on arrhythmia and ihd, insert stuff from first lit review on ecg methods.

TODO different methods of recording ECGs

TODO annotated ecg for parts, which diseases are apparent

TODO look at which ECG methods were covered in first lit review and take some of those

As mentioned above, SAX has already successfully been applied to ECGs, but there has not been much use of the method in that respect? TODO is that true?

MSAX has not yet been applied to ECG analysis.

1.2.1 ECG databases

TODO cite this shit

TODO turn this into at least one full paragraph

ECG data is patient data and thus not freely accessible in most cases. Online databases, most of them on Physionet, are an exception to this rule. Physionet provides databases on many types of medical data. They are all freely available and licensed for research and educational use.

REFERENCES

brockwell2016 [1] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, en, ser. Springer Texts in Statistics. Cham: Springer International Publishing, 2016, ISBN: 978-3-319-29852-8 978-3-319-29854-2. DOI: 10.1007/978-3-319-29854-2.

anacleto2020 [2] M. Anacleto, S. Vinga, and A. M. Carvalho, "MSAX: Multivariate Symbolic Aggregate Approximation for Time Series Classification," en, 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, ISBN: 978-3-030-63061-4. DOI: 10.1007/978-3-030-63061-4_9.

lin2003 [3] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," en, 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.

habozorgi2015 [4] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – A decade review," en, *Information Systems*, vol. 53, pp. 16–38, Oct. 2015, ISSN: 03064379. DOI: 10.1016/j.is.2015.04.007.

shieh2008 [5] J. Shieh and E. Keogh, "*I* SAX: Indexing and mining terabyte sized time series," en, in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, Las Vegas, Nevada, USA: ACM Press, 2008, p. 623, ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401966.

amahatana2005 [6] C. Ratanamahatana, E. Keogh, A. J. Bagnall, and S. Lonardi, "A Novel Bit Level Time Series Representation with Implication of Similarity Search and Clustering," en, in *Advances in Knowledge Discovery and Data Mining*, T. B. Ho, D. Cheung, and H. Liu, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2005, pp. 771–777, ISBN: 978-3-540-31935-1. DOI: 10.1007/11430919_90.

kaur2016 [7] I. Kaur, R. Rajni, and A. Marwaha, "ECG Signal Analysis and Arrhythmia Detection using Wavelet Transform," en, *Journal of The Institution of Engineers (India): Series B*, vol. 97, no. 4, pp. 499–507, Dec. 2016, ISSN: 2250-2106, 2250-2114. DOI: 10.1007/s40031-016-0247-3.

prasad2018 [8] B. V. P. Prasad and V. Parthasarathy, "Detection and classification of cardiovascular abnormalities using FFT based multi-objective genetic algorithm," en, *Biotechnology & Biotechnological Equipment*, vol. 32, no. 1, pp. 183–193, Jan. 2018, ISSN: 1310-2818, 1314-3530. DOI: 10.1080/13102818.2017.1389303.

panuccio2002 [9] A. Panuccio, M. Bicego, and V. Murino, "A Hidden Markov Model-Based Approach to Sequential Data Clustering," en, in *Structural, Syntactic, and Statistical Pattern Recognition*, G. Goos *et al.*, Eds., vol. 2396, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 734–743, ISBN: 978-3-540-44011-6 978-3-540-70659-5. DOI: 10.1007/3-540-70659-3_77.

corduas2008 [10] M. Corduas and D. Piccolo, "Time series clustering and classification by the autoregressive metric," en, *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 1860–1872, Jan. 2008, ISSN: 01679473. DOI: 10.1016/j.csda.2007.06.001.

nygaard1998 [11] R. Nygaard and D. Haugland, "Compressing ECG signals by piecewise polynomial approximation," en, in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 3, Seattle, WA, USA: IEEE, 1998, pp. 1809–1812, ISBN: 978-0-7803-4428-0. DOI: 10.1109/ICASSP.1998.681812.

zhu2018 [12] H. Zhu, Y. Pan, K.-T. Cheng, and R. Huan, "A lightweight piecewise linear synthesis method for standard 12-lead ECG signals based on adaptive region segmentation," en, *PLOS ONE*, vol. 13, no. 10, J. Zhao, Ed., e0206170, Oct. 2018, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0206170.

zifan2006 [13] A. Zifan, M. H. Moradi, S. Saberi, and F. Towhidkhah, "Automated Segmentation of ECG Signals using Piecewise Derivative Dynamic Time Warping," en, p. 5, 2006.

zan2016 [14] C. T. Zan and H. Yamana, "An improved symbolic aggregate approximation distance measure based on its statistical features," in *Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services*, ser. iiWAS '16, New York, NY, USA: Association for Computing Machinery, Nov. 2016, pp. 72–80, ISBN: 978-1-4503-4807-2. DOI: 10.1145/3011141.3011146.

sun2014 [15] Y. Sun *et al.*, "An improvement of symbolic aggregate approximation distance measure for time series," en, *Neurocomputing*, vol. 138, pp. 189–198, Aug. 2014, ISSN: 09252312. DOI: 10.1016/j.neucom.2014.01.045.

yu2019 [16] Y. Yu *et al.*, "A Novel Trend Symbolic Aggregate Approximation for Time Series," en, vol. abs/1905.00421, p. 9, 2019. [Online]. Available: http://arxiv.org/abs/1905.00421.

lkhagva2006 [17] B. Lkhagva, Y. Suzuki, and K. Kawagoe, "Extended SAX: Extension of Symbolic Aggregate Approximation for Financial Time Series Data Representation," en, in *Proceeding of IEICE the 17th Data Engineering Workshop*, Ginowan, Japan, 2006, p. 7. [Online]. Available: https://www.researchgate.net/publication/229046404_Extended_SAX_extension_of_symbolic_aggregate_approximation_for_financial_time_series_data_representation (visited on 02/27/2021).

alinowski2013 [18] S. Malinowski, T. Guyet, R. Quiniou, and R. Tavenard, "1d-SAX: A Novel Symbolic Representation for Time Series," en, in *Advances in Intelligent Data Analysis XII*, A. Tucker, F. Höppner, A. Siebes, and S. Swift, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2013, pp. 273–284, ISBN: 978-3-642-41398-8. DOI: 10.1007/978-3-642-41398-8_24.

fuad2010 [19] M. M. M. Fuad and P.-F. Marteau, "TOWARDS A FASTER SYMBOLIC AGGREGATE APPROXIMATION METHOD:" en, in *Proceedings of the 5th International Conference on Software and Data Technologies*, University of Piraeus, Greece: SciTePress - Science and and Technology Publications, 2010, pp. 305–310, ISBN: 978-989-8425-22-5 978-989-8425-23-2. DOI: 10.5220/0003006703050310.

camerra2010 [20] A. Camerra, T. Palpanas, J. Shieh, and E. Keogh, "iSAX 2.0: Indexing and Mining One Billion Time Series," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, Dec. 2010, pp. 58–67. DOI: 10.1109/ICDM.2010.124.

keogh2005 [21] E. Keogh, J. Lin, and A. Fu, "HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence," en, in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, USA: IEEE, 2005, pp. 226–233, ISBN: 978-0-7695-2278-4. DOI: 10.1109/ICDM.2005.79.

park2020 [22] H. Park and J.-Y. Jung, "SAX-ARM: Deviant event pattern discovery from multivariate time series using symbolic aggregate approximation and association rule mining," en, *Expert Systems with Applications*, vol. 141, p. 112 950, Mar. 2020, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2019.112950.

ordonez2008 [23] P. Ordóñez *et al.*, "Visualizing Multivariate Time Series Data to Detect Specific Medical Conditions," *AMIA Annual Symposium Proceedings*, vol. 2008, pp. 530–534, 2008, ISSN: 1942-597X. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656052/ (visited on 03/30/2021).

zhang2019 [24] C. Zhang *et al.*, "Anomaly detection in ECG based on trend symbolic aggregate approximation," en, *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2154–2167, 2019, ISSN: 1547-1063. DOI: 10.3934/mbe.2019105.

alghatrif2012 [25] M. AlGhatrif and J. Lindsay, "A brief review: History to understand fundamentals of electrocardiography," en, *Journal of Community Hospital Internal Medicine Perspectives*, vol. 2, no. 1, p. 14 383, Jan. 2012, ISSN: 2000-9666. DOI: 10.3402/jchimp.v2i1.14383.

fye1994 [26] W. Fye, "A History of the origin, evolution, and impact of electrocardiography," en, *The American Journal of Cardiology*, vol. 73, no. 13, pp. 937–949, May 1994, ISSN: 00029149. DOI: 10.1016/0002-9149(94)90135-X.

becker2006 [27] D. E. Becker, "Fundamentals of Electrocardiography Interpretation," *Anesthesia Progress*, vol. 53, no. 2, pp. 53–64, 2006, ISSN: 0003-3006. DOI: 10.2344/0003-3006(2006)53[53:FOEI]2.0.CO;2.

moody2001 [28] G. Moody and R. Mark, "The impact of the MIT-BIH Arrhythmia Database," en, *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, May-June/2001, ISSN: 07395175. DOI: 10.1109/51.932724.

oldberger2000 [29] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," en, *Circulation*, vol. 101, no. 23, Jun. 2000, ISSN: 0009-7322, 1524-4539. DOI: 10.1161/01.CIR.101.23.e215.

asilewski2012 [30] J. Wasilewski and L. Połoński, "An Introduction to ECG Interpretation," en, in *ECG Signal Processing, Classification and Interpretation*, A. Gacek and W. Pedrycz, Eds., London: Springer London, 2012, pp. 1–20, ISBN: 978-0-85729-867-6 978-0-85729-868-3. DOI: 10.1007/978-0-85729-868-3_1.

xie2020 [31] L. Xie *et al.*, "Computational Diagnostic Techniques for Electrocardiogram Signal Analysis," en, *Sensors*, vol. 20, no. 21, p. 6318, Nov. 2020, ISSN: 1424-8220. DOI: 10.3390/s20216318.