NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

THE LANGUAGE OF BIOSIGNALS

STRUCTURE AND MEANING IN TIME SERIES

JOÃO MANUEL DE ALMEIDA RODRIGUES

Doctoral/Ph.D dissertation

DOCTORATE IN NOVA INSTRUMENTATION FOR HEALTH (NI4H)

NOVA University Lisbon
December, 2022

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

THE LANGUAGE OF BIOSIGNALS

STRUCTURE AND MEANING IN TIME SERIES

JOÃO MANUEL DE ALMEIDA RODRIGUES

Doctoral/Ph.D dissertation

Adviser: Hugo Filipe Silveira Gamboa
Associate Professor with Aggregation, FCT-NOVA University Lisbon

Examination Committee

DOCTORATE IN NOVA INSTRUMENTATION FOR HEALTH (NI4H)
SPECIALIZATION IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
December, 2022

The Language of Biosignals

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.14) [112].

Dedicatory lorem ipsum.

Acknowledgements

This work is far from being only mine. It is the combined effort of all people and entities that followed me throughout this journey. Nothing here would have been possible without them. With all of you I shared doubts, frustrations, but also many victories. To all of you, I give my sincere aknowledgement.

I would like to start by thanking the *Faculty of Science and Technology from the NOVA University of Lisbon*, for giving me the environment and conditions to pursue this path, as well as to the financing entities, *Fundação para a Ciência e Tecnologias* and *Volkswagen Autoeuropa* for supporting this work. I also want to thank the Hanse-Wissenschaftskolleg for supporting my stay at Bremen and providing me an incredible experience and the Fullbright Comission for making my stay in Riverside possible.

A special thank you has to go to my family from the Biosignals Group. In this laboratory, I spent many challenging moments, but it was always a safe place where I could count on any of my dear fellows. I would start to acknowledge my mentor, prof. Hugo Gamboa, the main reason why I continued working in science. Although filled with doubts, prof. Hugo convinced me this was the right path. He was and always will be a reference, whom valued, inspired and supported me, but above all, challenged me in doing better and achieving more. Of the many challenges, some were not fulfilled and stayed as abstract ideas. But many others matured and bloomed into concrete works presented here in this manuscript. Prof. Hugo helped me grow as a researcher but also helped me shape part of who I am today. For all this, I am very grateful, hoping he enjoyed the ride as much as I did. David became a friend and older *brother*, showing support and companionship. He was essential in my early career. Thank you for all the advises, good conversations, and cheerful moments we spent together. Who thought that all would start by painting the lab furniture...To Cátia, central figure of the group, we shared so many good moments since the creation of the lab with so much *bricolage*. She taught me to make things easier and make better decisions, also being an important confident. Daniel, a source of knowledge, but mostly, a source of skepticism, highly necessary to challenge and test us in improving our work. I fondly remember our great trip to Denmark, thank you for everything. Ricardo, thank you for the energy and motivation that you shared so

nicely everyday you went to work. Nafiseh, thank you to make me know myself better. Phillip, thank you for all the wonderful conversations and discussions about so many topics (work and non-work related). Luís, such a motivating and energetic figure. Thank you for sharing so much knowledge, listening to my outbursts and giving me so many rides to Porto Brandão. To the master students I had the luck to follow or collaborate with, Guilherme, Diogo, António, Sara Santos and Sara Silva, Lua and Rita, thank you for teaching me so much. All the remaining members, Mariana and Daniel, I hope you can enjoy your time in the group and can continue the good work being done. I also want to thank our *old* neighbours, Paulo and Jorge. Thank you for your companionship, good mood and constant motivation. You were important pieces throughout these years.

I had the luck to share this journey with many, having collaborate with multiple entities. A special thank you goes to Fraunhofer-AICOS, specially to Duarte, for all the knowledge shared, your rigor, excellence, and work culture. Thank you also for all our enjoyable conversations about time series. Thank you to PLUX for all the equipment shared. Autoeuropa Volkswagen, thank you to the Industrial and Management Department, namely Pedro, Zé, Natércia, Jakie and specially Carlos, thank you for your availability and desire to make your ideas work. A big thank you goes to the Time Series group from the University of California, Riverside. Ryan, Audrey, Renjie and Maryam, as well as to Rutuja and Jiasi, thank you so much for the great environment you provided, my stay would not have been so pleasant without you. A special thank you goes to Eamonn Keogh. You are a great scientific inspiration. I always followed your work with a great interest and admiration. Thank you for receiving me so well, sharing so much knowledge and providing me such an amazing experience. I also had the chance to stay at the Cognitive Systems Lab from the Universität Bremen. It was a great pleasure and honour to work with such a wonderful group. Thank you to all the team members for receiving and treating me so well. A special thank you to Prof. Tanja, Felix, Rafael, Moritz, Yale and Hui. I will never forget the ideas, knowledge and great activities we shared. It was such a wonderful experience being at the CSL and I will forever cherish it. Thank you!

I kept the most dear to me for last. Many of them had to endure bits of my frustrations, but all helped to relieve it when gathering and enjoying our time together. Francisco, Nuno and Gonçalo, thank you for always being there when needed. My dear parents and sister, José, Angela and Nádia, you always guided me in the right way. I am sorry for being sometimes distant, and can't thank you enough for everything you did and always do for me. Ana, my dear wife, you lived and endured this as much as I did. I have to apologize for all the time I was *too busy* with something else. I can't thank you enough for your patience, moral support, and guidance, for joining me in this journey, and for simply being there when I needed the most, making everything easier. My achievements are yours too.

"Um ser humano é um ser que é constantemente 'em construção', mas também, de uma forma paralela, sempre em um estado de destruição constante" (José Saramago)

Abstract

As data scientists in the biomedical field, we strive to uncover the structure and underlying patterns within biosignals' data that allow gaining a clearer understanding of physiological phenomena. Methods and algorithms are used as means of interaction and *communication* to make the data *express* its content to retrieve the information we need. In the current data deluge with the advent of existing wearables and available biodevices, it is not always easy to unveil this meaningful information from biosignals. Hence, there is an emerging necessity to have tools that help analysts explore, analyse and retrieve this information and reduce the burden of such a time-consuming process.

In this thesis, this communication of the analyst using biosignals is explored answering the following questions: Can we *decipher* the *language of biosignals*? Can we then use words to describe specific patterns and shapes in biosignals? Are there any meaningful time-dependent strcutural changes in biosignals that can be *unveilled*?

Novel methods in time series analysis are proposed for fast, useful and meaningful information retrieval, to pave the way for more expressive and intuitive means of *communication*. More specifically, tools are proposed for (1) a practical way to automatically segment single-channel or multimodal biosignal data using a self-similarity matrix (SSM) computed with a signals' feature-based representation; (2) more expressive and intuitive pattern search strategies with (a) a novel symbolic representation of time series (SSTS), and (b) a word feature-vector used with operators (*Quo*TS); and (3) classification of time series, based on the mentioned symbolic approach (HeaRTS).

In topic (1) lucid visual support was delivered in interpreting biosignals with an SSM, while performing accurate automatic segmentation with the help of novelty and similarity functions, ultimately moving towards automatic labelling. As in (2.a), the novel symbolic representation was successful in deciphering the signal in several biosignals' use-cases, being used, such as (2.b) to perform pattern search with a simple vocabulary. These novel search mechanisms performed with good accuracy in detecting the desired patterns, but mostly with increased intuition and expressiveness. Finally, in (2.c) was showed that the signal could be translated into a text-based representation to perform classification task with standard text-mining classifiers.

Keywords: biosignals, time series, similarity, segmentation, meaning, structure, symbolic, text, expressiveness, classification, pattern, features

Resumo

Enquanto cientistas de dados no ramo da engenharia biomédica, procuramos desvendar estruturas e padrões, inerentes aos biosinais, que permitem compreender fenómenos fisiológicos associados. Métodos e algoritmos são utilizados como meios de interação e comunicação de forma a permitir aos dados expressarem o seu conteúdo para extrair a informação de interesse. Nem sempre se torna simples retirar esta informação, sendo que, especialmente com o uso mais frequente dos biodispositivos *vestíveis*, torna-se cada vez mais relevante ter acesso a ferramentas que ajudem os analistas a explorar, analisar, e recolher informação, para reduzir a dificuldade, tempo e custo deste inevitável processo.

Neste trabalho, a comunicação entre o analista e os biosinais é explorada, procurando responder às seguintes questões: Podemos *decifrar* a *linguagem dos biosinais*? Podemos usar palavras para descrever padrões e formas específicas dos biosinais? Haverão mudanças estruturais que podem ser desvendadas?

Novos métodos são propostos para a análise mais rápida, útil e significativa das séries temporais, abrindo o caminho para um meio de comunicação mais expressivo e intuitivo entre o analista e os dados recolhidos. Mais especificamente, são propostas ferramentas para (1) a segmentação automática de biosinais uni ou multimodais, por meio de uma matriz de auto-similaridade (SSM), calculada a partir da representação das características do sinal; (2) procuras de padrões mais expressivas e intuitivas com (a) uma nova representação simbólica das séries temporais (SSTS), (b) palavras que representam características significativas usadas com operadores (*QuoTS*); e (3) classificação de séries temporais com base na representação simbólica dos sinais (HearTS).

Em relação ao tópico (1) foi apresentado um suporte visual lúcido para interpretar os biosinais com a SSM, enquanto também é possível realizar a segmentação automático com o uso da função de novidade e similaridade, a caminho do processo de categorização automática. Em (2.a), a representação simbólica proposta mostra-se competente em diversos casos de uso de biosnais, tal como o método proposto em (2.b), no contexto de procura de padrões em biosinais, com a ajuda de um vocabulário simples. Estes dois métodos propostos apresentaram resultados sólidos na idenficação dos padrões desejados, sobretudo com mais intuição e expressivadade. Finalmente, em (2.c) foi apresentado que

o sinal poderia ser transformado em texto e ser usado em problemas de classificação utilizando classificadores tradicionalmente usados em texto.

Palavras-chave: biosinais, séries temporais, similaridade, segmentação, significado, estrutura, simbólico, texto, expressividade, classificação, padrão, características

Contents

List of Figures

List of Tables

Acronyms

ABP Arterial Blood Pressure *(pp. xiii, xv, xvi, xx, 15, 20, 21, 43, 55, 57, 58, 65–67, 69, 97, 110, 111, 114, 136)*

ACC Accelerometer *(pp. xiv, xviii, xix, 43–45, 50, 51, 75, 97, 106, 107)*

BoW Bag of Words *(pp. xvii, 24, 25, 79, 80, 82, 84, 85, 119–121, 124, 137)*

DE Duration Error *(pp. xxiii, 103)*

DFT Discrete Fourier Transform *(pp. xiii, 13–15)*

DoF Domain of Freedom *(p. 51)*

DTW Dynamic Time Warping *(pp. xiii, 16–18, 71)*

EAWS Ergonomic Assessment Worksheet *(pp. 4, 138–140)*

ECG Electrocardiogram *(pp. x, xiii, xiv, xvi, xx, 2–4, 12, 17, 18, 22, 34, 43, 47–49, 55, 65, 67, 97, 99, 110, 128–130, 136, 138)*

ED Euclidean Distance *(pp. xiii, xvii, 16–20, 22, 23, 62, 71, 84, 85, 91, 97, 119–121, 138)*

EEG Electroencephalogram *(pp. 3, 34, 43)*

EMG Electromyogram *(pp. xiv, 3, 4, 34, 43, 47, 50, 97)*

EOG Electrocularogram *(p. 43)*

F1 F1-measure *(pp. xxiv, 186)*

fNIRS functional Near InfraRed Spectroscopy *(p. 4)*

GON Goniometer *(p. 50)*

GYR Gyroscope *(pp. xiv, xviii, xix, 43–45, 50, 51, 97, 106, 107)*

HAR Human Activity Recognition *(pp. 43, 57, 63, 97, 99)*

HeaRTS Human Readable Time Series *(pp. 74, 97, 119, 121, 135, 137, 140, 146, 147)*

idf inverse document frequency *(p. 84)*
IMU Inertial Motion Unit *(pp. 3, 4, 50, 51)*

LDA Latent Dirichlet Allocation *(p. 147)*
LSA Latent Semantic Analysis *(p. 147)*

MAE Mean Absolute Error *(p. 28)*
MAG Magnetometer *(p. 51)*

NLP Natural Language Processing *(pp. 134, 137, 146)*
NMF Non-Negative Matrix Factorization *(p. 147)*
NNbr Nearest Neighbor *(p. 23)*

OCRA OCcupational Repetitive Action *(p. 4)*
oprd operands *(pp. 29, 116)*
oprt operators *(pp. 29, 116)*

P Precision *(pp. xxiv, 186)*
PAA Piecewise Aggregate Approximation *(pp. xiii, 15, 20, 23)*
PrevOccupAI Prevention of Occupational Disorders in the Public Administration with AI *(p. 139)*
PYTS Python for Time Series Classification *(pp. xiii, 15, 16)*

QuoTS Query on Time Series *(pp. xviii, xx, 49, 73, 74, 86, 88, 92, 97, 125–130, 132, 133, 135, 137, 138, 143, 147)*

R Recall *(pp. xxiv, 186)*
regex regular expressions *(pp. xvi, xvii, xxii, 23, 25, 26, 28, 73, 74, 77–79, 81, 83, 86, 93, 94, 109, 117–119, 135–137, 146)*
RESP Respiratory Inductance Pletismography *(p. 65)*
RULA Rapid Upper Limb Assessment *(pp. 4, 139)*

SAX Symbolic Aggregate Approximation *(pp. xiii, 15, 16, 23, 74, 137)*
SCR Skin Conductance Response *(pp. 49, 130)*
sf similarity function *(p. 102)*
SNR Signal-to-Noise-Ratio *(p. 99)*
SSM Self Similairty Matrix *(pp. xii, xiv–xvi, xviii, xix, xxiii, 22, 33, 35, 37, 39, 53, 54, 56–65, 67, 69, 70, 72, 96, 97, 101–109, 134–137, 139, 142–145, 165)*

SSTS Syntactic Search on Time Series *(pp. xii, xvi, xvii, xx, xxii, xxiii, 73–76, 78–82, 85, 86, 96, 97, 109, 114–116, 118–120, 125, 135, 137–139, 146, 147)*

SVD Singular Value Decomposition *(p. 13)*

SVM Support Vector Machines *(pp. xx, 25, 119–122, 124, 125)*

tf Term Frequency *(pp. xvii, 24, 82, 84)*

TF-idf Term Frequency - inverse Document Frequency *(pp. xvii, xx, 24, 25, 79, 80, 82, 84–86, 97, 119–122, 124, 137, 138)*

Toprd total number of operands *(pp. 29, 116)*

Toprt total number of operators *(pp. 29, 116)*

VOC Volatile Organic Compounds *(p. 140)*

WMSD Work Related Musculoskeletal Disorder *(p. 3)*

1 Introduction

1.1 Biosignals and Challenges of a Data-Driven Society

It has never been easier to gather information about any aspect of our life, work, health, industry, education, and society in general. The continuous collection of data through the usage of mobile phones, smartwatches, hearables, wristbands, and other non-invasive wearable sensors leads to a large and valuable volume of information gathered in all possible scenarios. As reported in *Tankovska et al.*, the use of wearable devices has more than doubled in the interval between 2016 and 2019, reaching 722 million worldwide [183]. This data often comes in the form of time series, which is one of the most common data types in nature [79].

A time series data type with special interest is a *biosignal*. These are time series generated by the *human body*, such as the *heart* (ECG - ), *muscles* (EMG - ), *brain* (Electroencephalogram (EEG) - ) or even movements (Inertial Motion Unit (IMU) - ). The usage of such signals has emerged as Biosignal-based technology has been increasingly available in our daily lives, being widely applied in biometrics [21], human activity recognition [34, 36, 8], sports sciences [101, 109, 128, 87, 75, 127, 74, 211, 46, 144], healthcare [206, 181, 121, 25, 131, 19], rehabilitation assistance [45, 23], ergonomics [168, 167], and edutainment [95].

Having relevant information about a specific topic can be beneficial, but the overwhelming amount of data being collected brings tremendous challenges in the ability to save, process, and analyze it, as well as retrieve interpretable and meaningful information based on which we can act upon [184]. Ultimately, these challenges make it difficult to have data that is well structured and labelled, considering that it is an intricate, thorough, and time-consuming process. Additionally, higher quantity of data increases the complexity of the mentioned processes. In the work of *Roh et al.* it is mentioned that data scientists only rely on a small portion of the available datasets as it is too expensive to label all the data available [163]. This is just an example of how much data can have limited use or even go unused. This is particularly problematic when developing machine learning applications, as data should be correctly pre-processed and labelled for supervised approaches so it

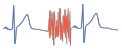
does not include noise, artifacts or mislabeled segments of the signal [163].

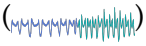
Computational methods are necessary to reduce time-consuming data-driven analysis, but the first layer of knowledge comes from the human mind and not from machines. Taking this into account, interactive data analysis is a great challenge as there is a lack of integrated interactive systems, tools and methods for fast, useful, and meaningful information retrieval from time series data [71, 72, 70, 118]. In order to do more with the data we have, tools should be available to support and help analysts to accelerate the process of information retrieval from time series. We believe that having more informative, expressive, and intuitive methods for the analysis of such data can help researchers that are familiar with time series data mining, such as data scientists, in accelerating their analysis tasks and streamline their analysis tasks. In addition, we also believe that such methods could help democratize these tasks to researchers that are not as experienced in this domain, but could benefit from it in their research activities [99].

In this thesis, we explore two main domains of methods for information retrieval in time series, which will be referred to as (1) *unveiling the grammar of time series* and (2) *a language for time series data mining*. In the first domain, we propose a practical and manageable way to automatically segment single channel or multimodal time series. The proposed approach also brings a lucid visual support in interpreting the time series and helps *unveiling* its structure. The second domain focuses on using text-based approaches to search for patterns and events on time series with expressive queries, moving towards human interpretable and readable time series analysis. A novel symbolic representation of time series will be proposed and its application on query-based search and classification of time series is explored.

The presented thesis is entitled *A Language for Biosignals*. It suggests that concepts and algorithms used for text in natural language processing can be applied to *biosignals*. Thus, we will further explain what we believe is the *linguistic nature* of time series, with a deeper focus on *biosignals*.

1.2 Linguistic Nature of Biosignals

The human brain has an innate ability to identify relevant structures and patterns from the information it receives, being fundamental for decision-making [137]. This capacity is also revealed when observing time series, and finding inherent patterns that can be associated with specific phenomenons. As an example, let us depict an ECG, a typical physiological signal. The standard cyclic nature of the ECG may be affected by several sources, such as motion artefacts, muscular contractions or even symptomatic events. For instance, the signal piece  has two cycles of an ECG disturbed by noise. Thus we can conclude that the signal has three segments, of which two of these (the first segment and the last) have high similarity, i.e., *A B A*. Other changes may be related with the dynamics of the process, for instance, a subject can change and evolve his/her gait from

walking to jogging () — this ACC signal contains two main periodic *regimes*, which could be segmented as WW...W and JJ...J.

This visual intuition is also straightforward when an analyst is searching for specific shapes or patterns in time series. Often, the analyst resorts to describe the shape he/she is looking for. For instance, a physician may say "*I am searching for the T-wave, that represents the* `wide peak`" () or "*I am searching for the QRS complex, that looks like a* `sharp peak followed by a sharp valley` ()". This visual intuition also happens when analysts are trying to find differences between classes of signals. For instance, the following shapes (1)  and (2)  are different because "*shape 1 has a* `positive peak` *where shape 2 doesn't*".

Time series are carriers of information, and the presence of changes in its *regimes* or the presence of specific shapes in its segment may be associated with an occurrence in the physical world attributed to some meaning. These underlying associations that the human mind does give this notion of structure and meaning, which is a good approximation of what represents the foundation of a language: grammar and meaning [39]. In language, a sequence of words needs to follow a certain structure to have an interpretable meaning. In biosignals this can also be observed as the way how certain patterns are structured in the signals and allow us to extract a physiological meaning. Thus, we can pose the question: Do biosignals exhibit a certain grammar and meaning, just like language does?

Grammar is generally defined as the set of rules that constitutes the structure of a language and is modeled by morphology and syntax [39]. The first is the structure of words, how these are built or morphed from individual symbols based on context, while the latter consists in organizing words in sequences to form larger linguistic units, such as sentences. Just as a language has morphological and syntax rules that represent its structural information, time series can be organized by a formal structure of ordered sub-segments with specific morphological characteristics, organized to build larger segments. Our first topic is related to *unveiling* this structure with methods that can parse it.

In addition to a *grammar*, a language also has *meaning*. The *meaning* regarding time series depends on the context and what occurred in the physical world that is then seen on the signal. Specific occurrences might be attributed a specific meaning by an analyst, as we have seen above with the physician example. In this work, we explore a language to *translate* time series into text and use this textual information as an expressive way of searching for meaningful events and patterns. This is related to our second topic, which focuses on using language in time series data mining tasks.

Having explained how time series have a *linguistic nature*, we now focus our attention to a specific domain of *biosignals* and how the questions that are generally posed on time series are also valid, meaningful, and applicable to *biosignals* from the occupational health context.

1.3 Biosignals: Context and Relevance in Occupational Health

This thesis was developed in partnership with the *Ergonomics* team from *Volkswagen Autoeuropa*, with a focus on the analysis of *biosignals* from workers to retrieve meaningful information about their occupational risk status and prevent Work Related Musculoskeletal Disorder (WMSD). WMSDs prevail as the most common occupational diseases in the European Union and have a global impact on the well-being of individuals and their quality of life in a range of working sectors [83], accounting for the second-largest cause of disability worldwide [116]. These are especially prevalent as upper limb or neck disorder (with 42% of all WMSD cases reported) [62] in several industry sectors, such as textile and automotive, where production processes with pre-defined motions and actions have a repetitive nature. This type of occupational activities have a negative impact on the well being, and making workers more prone to such disorders, with tremendous consequences to both workers and companies, namely absenteeism, early retirement, and loss of productivity [189, 190].

Several strategies have been implemented to identify, regulate and prevent occupational risks in manufacturing industries, such as (1) the inclusion of job rotation schedules, which promote a variation of the exposure throughout the working day [12, 160] and (2) screening tools, for the assessment of occupational risk exposure, for example the OCcupational Repetitive Action (OCRA), Rapid Upper Limb Assessment (RULA) or the Ergonomic Assessment Worksheet (EAWS) [143, 126, 172]. Nevertheless, these strategies are not optimal because they (1) are not automated, relying on observational methods and dedicated personnel to inspect video records; (2) are not objective measures; (3) do not take into account differences among the worker's population, such as anthropometric parameters, age and experience variability; and (4) present single scores, being insufficient to explain the factors that contributed to this risk. With the advent of Industry 4.0, more companies are using modern strategies that follow digital solutions to provide direct and objective quantitative measures [164]. An example of these strategies is the usage of *biosignals*, with wearable inertial devices for physiological, motion, and posture tracking of workers.

From IMU, time series can be collected and relevant variables and parameters can be directly measured, such as the position and velocity of each body segment, postural angles between joints, and gait parameters, making these important for ergonomics studies [27, 195]. There are some limitations to using IMU, mostly related to the long-term bias (sensor drifting) arising from long acquisitions and the empirical process to fine-tune sensor fusion techniques. Other systems can be used for motion capture, such as camera-based methods, but these rely on a fixed setup of cameras, which is unfeasible in real industrial scenarios [167]. In addition to motion sensors, the inclusion of physiological sensors, such as ECG, EMG and even functional Near InfraRed Spectroscopy (fNIRS) can give reliable evidence of other occupational health variables, namely cardiovascular load, muscular activation, cognitive effort and fatigue [178, 98, 69, 55].

The usage of biosignals in this context can play an important role in supporting the decision of ergonomists and other professionals in the industry. In order to develop systems that can use physiological, motion, and postural data for direct risk assessment and reporting, several challenges have to be answered. For instance, considering the periodic nature of most manufacturing tasks, risk factors are calculated by working cycle. Therefore, methods should be developed to identify working cycles. In addition, real occupational scenarios might have interruptions or changes in the working behaviour, due to abrupt production stoppage, shift breaks, or even because the worker shifted to another workspace that has a different movement pattern.

Other questions brought up by ergonomists, such as *can we find a pattern that has a sharp rise in the acceleration of the arm motion?* or *when the worker is using a hand tool to rotate a screw, can we see a periodic pattern on the acceleration pattern from the hand?*, which represent specific patterns with a descriptive shape that can be seen on the signals and are specific of a task, need to be explored qualitatively and quantitatively. These events can be relevant to studying their precise impact on the worker's occupational exposure. Having ways to detect these patterns is of great relevance as well. In this study, we will show how the proposed solutions can have an impact on these problems, and how they contribute to providing relevant visual feedback for information retrieval and make the search for specific patterns more intuitive and expressive, even profiting from the expressiveness of the text written on the questions above.

1.4 Research Paths

The central objective of this work is the development of novel tools for biosignals information retrieval. Overall, the work in this thesis contributed to all layers related to *biosignals*, from the moment data is acquired (*sensing*), to when it is processed for information retrieval (*analysis*), and how this information is used to act upon (*decision-making*). These three topics were the starting point of this work. It eventually branched into several research paths, specially for the *analysis* layer. A detailed description of each of the researched layers is provided further:

1. **Sensing** - Explore in depth the available technology to measure motion and postural variables in occupational scenarios for risk assessment. This will take into account which variables are associated with a risk, based on ergonomic standards. These measures are returned as *biosignals*, which are processed in the topic *analysis*;

2. **Analysis** - In this topic, two main research paths are explored:

    a) ***Unveiling the Grammar of Time Series*** **-** Research how to perform structural information retrieval in *biosignals* with a special interest in the segmentation based on *novelty* and *periodic* search, with applications to *summarization*, and how the resulting segments are related to each other, based on their *similarity*.

For this, we will introduce the **SSM**, a feature-based transformation of the signal from similarity-based measures that makes a meaningful visual representation, from which the segmentation points can be extracted and the relationship between segments can be made.

b) ***A Language for Time Series Data Mining -*** Explore a symbolic representation of *biosignals* and a word feature-based representation of *biosignals*, studying how these can be used for more expressive and intuitive pattern search with the help of regular expressions, natural language and operators. In this research branch, we will introduce SSTS and QuoTS.

From the textual representation of time series, we also explored how to make a higher level distance measure, following standard text mining methods. The resulting outputs of these methods can ultimately be used to have more hints of why a signal is different from another. This path will introduce HeaRTS.

3. **Decision-Making -** Discuss how the developed methods can contribute to more aware and informed decisions. Considering the outputs of the methods developed in research path **2.a - Analysis - Unveiling the Grammar of Time Series**, how can the analyst gain intuition over the structure of the data associating it with what happened in the physical world? With regards to research path **2.b - Analysis - A Language for Time Series Data Mining**, how expressive is the process of searching for specific events and patterns with the proposed linguistic-based search methods?

1.5 Thesis Structure

This thesis provides a detailed description and explanation of the research work developed during the Ph.D. program. It is organized into seven main Chapters and two Appendixes, each emphasizing on the scientific contributions developed throughout this thesis. Figure 1.1 illustrates of the structure of this work with each Chapter's topics and content. **Chapter 1 - Introduction -** has introduced the main motivations, goals, and context for the development of the proposed methods. **Chapter 2 - Time Series Fundamentals -** explains the fundamental definitions and knowledge needed to have a clear picture of what is developed in this work. **Chapter 3 - State-of-the-art -** depicts the literature review of related work, namely in the topics of segmentation, summarization, pattern/event search, and classification, with a focus in *biosignals*. **Chapter 4 - Unveiling the grammar of time series -** explains the algorithm developed for time series structural information retrieval and provides examples of its usage for novelty segmentation, periodic segmentation, similarity profiles, query-based search and summarization. **Chapter 5 - A language for time series data mining -** covers the usage of language for time series data mining, more specifically introducing a novel symbolic approximation and how it can be used for pattern search and classification. In addition, it also explains how to use natural language with a word-feature-based representation of time series. **Chapter 6 - Experimental Analysis,**

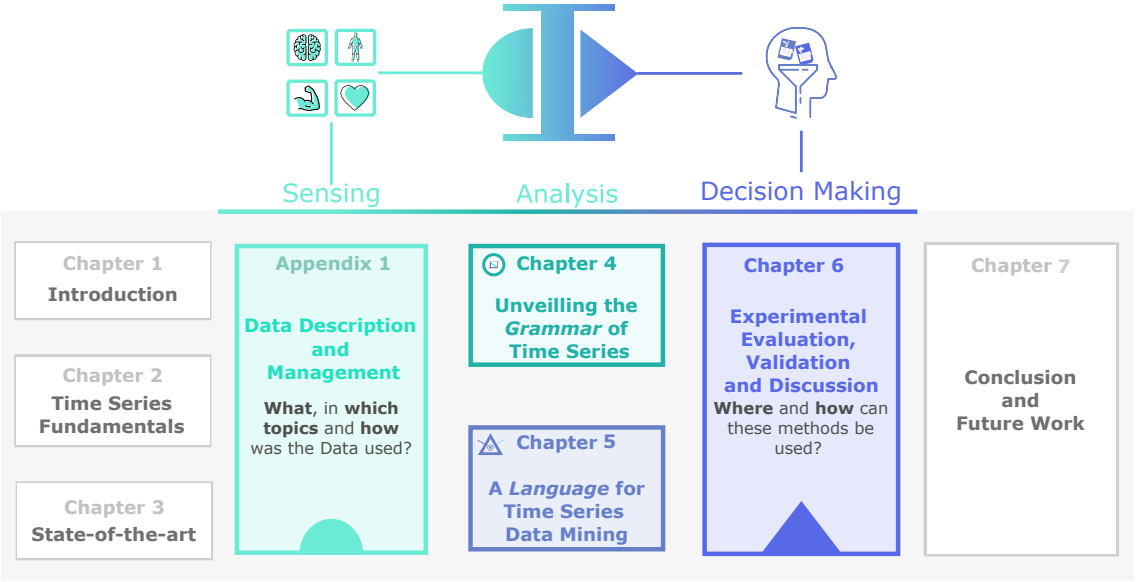


Figure 1.1: General topics on this thesis and structure of the document. It highlights the 3 layers of involvement related to time series: Sensing, Analysis, and Decision-Making, focusing on the Analysis layer, which includes two major topics subdivided into 4 sections each.

Validation and Discussion - shows the application of the previous methods to a set of examples, and presents major results that validate the performance of the proposed methods. In addition, this chapter also provides a general discussion of the results and how the proposed methods can be used for the benefit of the analyst. **Chapter 7 - Conclusion and Future Work -** gives an overall remark on the outcomes of this thesis and a reflection on the contributions that the developed methods have in making time series preparation and data mining more expressive, quicker, and more practical for an ever-increasing number of data available. It also gives a clear idea of which are the future paths for this work in terms of novel applications and performance improvement.

Two Appendixes are also part of this work and give additional information on the datasets used, algorithms and results. **Appendix 1 - Data description and management -** describes the data we used, explaining its source for both private data and publicly available data, for which purposes it was used and how it was used in this work. **Appendix 2 - Detailed Results for Information Retrieval and Classification -** gives more non-fundamental information about the methods developed, and detailed results, namely of the algorithm's performance in each record of the datasets, and hyperparameters used.

2

Time Series Fundamentals

The content of this thesis is diverse and covers several different topics. Therefore, we will start by setting the foundations that are necessary to fully capture the essence of this work. For this, we introduce the concepts and topics covered, and provide the necessary definitions and corresponding notation used in this work. Let us begin by explaining global definitions related to *time series*, which is the data type of interest in this work. Then, examples of *time series*, in particular *biosignals*, are given. Further, standard pre-processing methods, representation forms, and distance measures are also explained.

Additionally, as this work makes a strong connection between time series and their textual nature, the association between text and time series is introduced in this chapter. Finally, we also explain the standard quantitative metrics used to show the results and validate the proposed methods.

2.1 Global Definitions

The information gathered by sensors is a physical quantity that varies with time. These are called *time series* and are the main topic of this work.

Time Series: A *time series* is a sequence of real values ordered in time with length $n \in \mathbb{N}$: $T = (t_1, t_2, ..., t_n)$. A *biosignal* is a category of *time series*. Several data domains rely on the *multidimensional time series* acquisition from one sensor's multiple axes, such as an accelerometer's three directions, or multiple sources, such as an IMU that fuses three different sensors.

Multidimensional Time Series: A *multidimensional time series* is a set of $k \in \mathbb{N}$ *time series* belonging to the same acquisition: $\{T_1, T_2, ..., T_k\}$. Segments of interest, called *subsequences*, are often searched inside a *time series*.

Subsequence: A *subsequence* is a segment of a (*multidimensional*) *time series* with size $w \in \mathbb{N}$, starting from a given position i and ending at position $i+w$. Therefore, two instants,

defined as *events*, delimit a *subsequence* in time.

Event: An *event* is an instant in time e that indicates the presence of a relevant occurrence in the *time series*. Multiple *events* segment the time series into several *subsequences* of different lengths. Hence, *event* detection is often considered *time series* segmentation or change point detection[26]. To be clear, we will use the terms *event detection* and *segmentation* when discussing our methods, but can eventually use the term change point detection when comparing with other methods. A common strategy used in time series data mining to find relevant *subsequences* or *events* is the *moving window*.

Moving Window: A *moving window* is a process of sliding along a time series T to apply a specific method on each *subsequence* it hovers, a common strategy used in *time series* data mining to find relevant *subsequences* and *events*. The window has, similar to the *subsequence*, a predefined size $w \in \mathbb{N}$, which starts at a given position i and ends at position i+w. The process of *moving windows* is iterative, and windows can overlap each other. The next window will start at i+o, where $o \in [1, w]$ is the overlapping size ($o = 1$ for total overlap and $o = w$ for no overlap). On each *moving window* from each *subsequence* of the (*multidimensional*) *time series*, features can be extracted to form a *feature series*.

With a moving window, each *subsequence* of a *time series* can be filtered, features can be extracted or distances can be measured. We will show several utilities of this technique further when introducing methods used to pre-process a raw time series and apply standard distance measures. More importantly, we start by explaining the importance of a moving window to represent a *time series* with a feature, a *feature series*.

Feature Series: A *feature series* is a feature representation of a time series with size $m = \frac{n}{w-o}$ that depends on the overlap size $o \in \mathbb{N}$ of the *moving window*. In the case of a *multidimensional time series*, the *feature series* stack a *multifeature series* with size $f_{k,m}$. Multiple features extracted from one dimension or various dimensions are grouped into a *feature matrix*.

Feature Matrix (F_M): A *feature matrix* with size $r \times (k \times m)$, represents that each of the k dimensions produces r features. This *feature matrix*, which characterizes the (*multidimensional*) *time series* in statistical, temporal or spectral domains, is used to compute the *self-similarity matrix*.

Having introduced essential concepts for the understanding of the next few sections, we further give an overview of essential techniques used in time series processing and analysis, namely filtering methods, representation methods, distance measures and its applications, and a formal explanation of the relationship between text and *time series*.

2.2 Filtering

Time series have multiple sources of disturbance. This disturbance is usually called *noise* and is defined as an unwanted form of energy, but it can have multiple interpretations. It can be caused by internal sources inside a device, such as *white noise*, or be due to external sources, such as motion artifacts, wandering baseline, sensor detachment, or the magnetic field from surrounding devices. Any of these disturbances will affect the analysis stage and should be detected or removed.

2.2.1 Spectral Filtering

Several methods can be used to reduce the influence of noise in the analysis. Standard filtering methods, such as low-pass, band-pass, and high-pass filters can be used to reduce the presence of specific frequency bandwidths that are not relevant. There are many configurations for these types of filters, being one commonly used the *Butterworth* filter, with the following frequency response:

$$H_{j\omega} = \frac{1}{\sqrt{1 + \epsilon^2 \left(\frac{\omega}{\omega_c}\right)^2}}, \tag{2.1}$$

where n is the order of the filter, ω the frequency ($\omega = 2\pi f$), and ϵ the maximum amplitude gain.

2.2.2 Smoothing

Another method, often used to reduce the presence of noise and is a variant of a low-pass filter, is the smoothing technique. Several variations of this method exist, being the simplest one a moving average, which uses a moving window to calculate the mean in each iteration. Other approaches also convolve the signal with a specific window (H) (e.g. *Hanning window*), which instead of giving the same weights to all the samples of the moving window (moving average), attributes a higher weight to the moving window's central samples.

$$Tm_i = \Sigma_{j=a}^{a+w} T_j H_{j-a}, \tag{2.2}$$

where Tm_i is the i^{th} smoothed sample of the time series T, segmented by a and $a + w$ (w is the size of the moving window) and H is the window function used to smooth the signal.

2.2.3 Wandering Baseline

Another type of disturbance on the data that is usually removed is a wandering baseline. An example typically occurs in ECG signals, where the respiration creates a wandering baseline on it. This type of disturbance has a very low frequency compared to the

meaningful information on the data and can be removed by subtracting a *smoothed* version of the original data or applying a high pass filter.

2.3 Normalization

Normalization of data is an important step in any data mining process. It is essential for data standardization and scaling while keeping the morphology and shape of the time series. Several methods can be used for this purpose, namely:

$$\overline{T} = \frac{T}{max(|T|)}, \tag{2.3}$$

where the normalized signal ($\overline{T}$) is scaled by the absolute maximum of T. It is the simplest approach to normalization and guarantees that values are scaled linearly and their modulus cannot be higher than 1.

A variation of this process is the normalization by the range of amplitudes:

$$\overline{T} = \frac{T - min(T)}{max(T) - min(T)}, \tag{2.4}$$

where the time series T is normalized to range between [0,1]. Another normalization method, called *z-normalization*, is very commonly used and relies on the distribution of its values:

$$\overline{T} = \frac{T - \mu_T}{\sigma_T}, \tag{2.5}$$

where the time series T is subtracted by its mean, μ_T and scaled by its standard deviation, σ_T. The resulting values represent how many standard deviations the signal is away from the mean.

2.4 Transformation

In information retrieval, data has often to be re-scaled, simplified, approximate, or represented into another data type. Each can contribute in their way to capture the most relevant and meaningful information, or discover a new type of information that once was hidden in the original data. Dozens of methods exist for time series representation, such as Singular Value Decomposition (SVD) or wavelet transform, but only the ones relevant to this thesis will be explained.

2.4.1 Spectral Transformation

One of the first and most well-known techniques suggested for time series transformation is the DFT [3]. The idea behind this concept is that any signal, either periodic or not, is a decomposition of a finite number of sine and cosine waves, from which amplitude and phase are used to represent the signal in the frequency domain [177]. This transformation

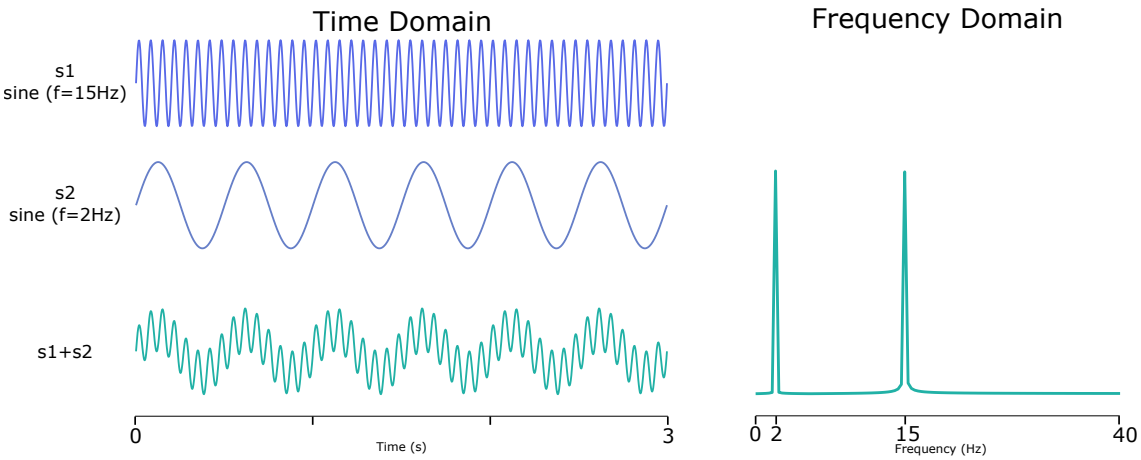


Figure 2.1: DFT of a sum of sine waves.

allows us to see the signal differently, highlighting which frequencies concentrate more or less energy. It unveils the presence of specific types of noise or artifacts, or periodic shapes. Figure 2.1 shows the transformation of a signal into the frequency domain. The signal is the sum of two different sine waves with 2 and 15 Hz respectively. The result of the Fourier transform is a frequency series with high energy at the frequencies of the sine waves.

The frequency domain opens new possibilities for the analysis of the time series, namely for the extraction of spectral features, which are also used to represent the signal in the feature domain.

2.4.2 Feature-based Representation

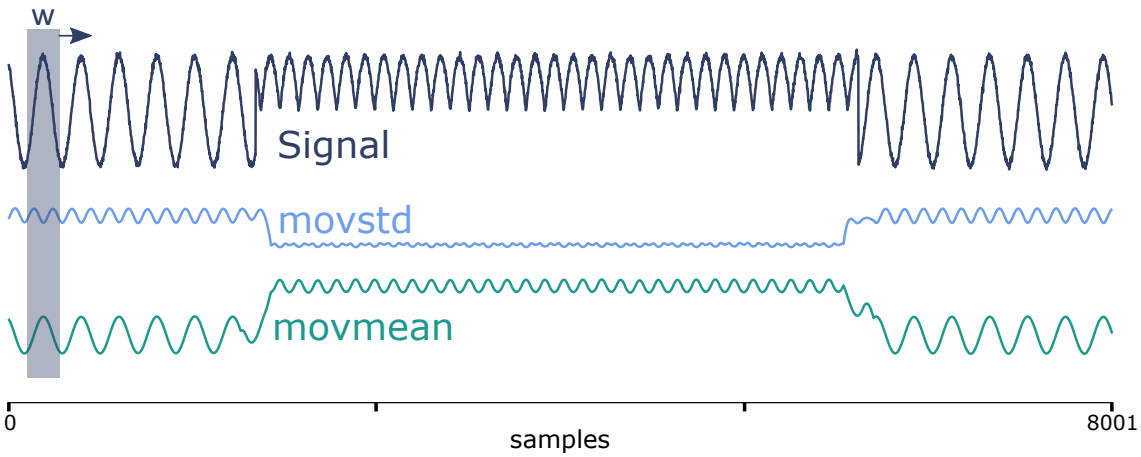


Figure 2.2: Moving window used to extract features with total overlap. The mean and standard deviation are extracted from the signal. The left shows the sine waves, while the right shows the frequency spectrum of the combination of sine waves.

The process of feature extraction is also a transformation method commonly employed and is performed by a *moving window*, resulting, for each feature, in a feature series (see

Section 2.1). In Figure 2.2 is shown a time series from which the average (moving mean) and standard deviation (moving *std*) are computed with a moving widow of size $w = 100$ and a total overlap $o = 100\%$. The feature is a characterization of the signal for that specific extracted property and is commonly used in machine learning tasks. In this thesis we will present two main applications of the feature representation of time series.

2.4.3 Piecewise Aggregate Approximation

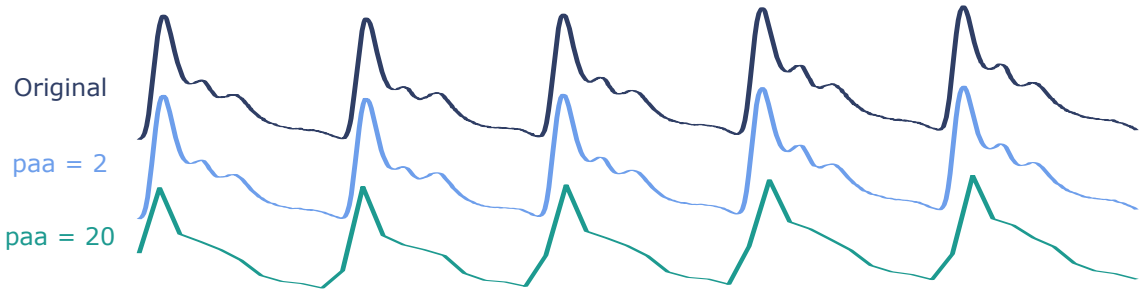


Figure 2.3: PAA representation of a ABP signal, with window sizes of 2 and 20, repsectively. The PAA representation was computed with the PYTS module [47], based on [93].

Another common used transformation method to simplify a time series and reduce its dimension is the PAA) [93]. The new representation space will have size $1 < N \leq n$, in which N is a factor of the original size n. The method searches to keep the average of the N equi-sized *subsequences* in which the original signal with length n is segmented, which results in $\overline{T} = \overline{t_1}, \overline{t_2}, ..., \overline{t_N}$, such that [93]:

$$\overline{t_i} = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} t_j. \tag{2.6}$$

An example is shown in Figure 2.3, where a ABP signal is transformed into a PAA with sizes 2 and 20, respectively.

2.4.4 Symbolic Aggregate Approximation

From this method, a new representation technique was born, transforming the signal from the numerical to the symbolic domain. It is called SAX [103]. This method applies PAA to a z-normalized time series and indexes a *character* to each sample of the simplified signal based on the distribution of its amplitude values [103]. The signal's amplitude values are separated in bins with equal probability. The number of bins is equal to the size of the symbolic *alphabet* chosen. Figure 2.4 shows an example of the signal transformed into a string with 3 letters in its alphabet. Such as the DFT, SAX brings novel ways of analising time series in a completely different manner, profiting from the much-acquired knowledge in text mining.

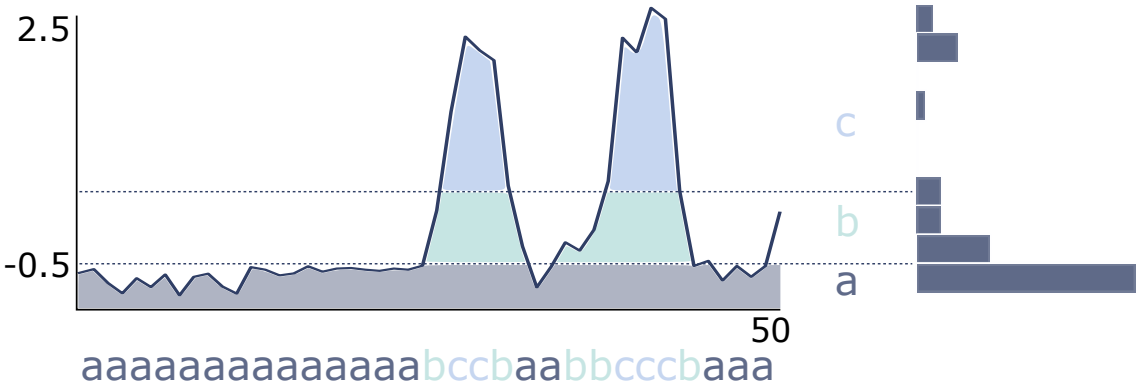


Figure 2.4: SAX representation of a power consumption signal from a Dutch Company, with window bin size of 3. The SAX representation was computed with PYTS based on [103].

This representation method is relevant to be introduced as it was a strong inspiration for the proposed novel symbolic representation technique for time series. Ultimately used for expressive pattern search and classification.

In order to perform pattern search or classification tasks, we have to calculate the difference/similarity between two time series or *subsequences*.

2.5 Distance Measures

There is a large number of distance measures for time series, but two of the classical and standard measures still provide state-of-the-art results in most time series data mining tasks, namely the ED and the DTW.

2.5.1 Euclidean Distance

The ED is the most straightforward distance measure for time series. Let us consider two time series, Q and C, of length n, so that

$Q = q_1, q_2, ..., q_i, ..., q_n$
$C = c_1, c_2, ..., c_i, ..., c_n$

The distance between these two time series under the ED is:

$$ED(Q, C) = \sqrt{\Sigma_{i=1}^{n}(q_i - c_i)^2}, \tag{2.7}$$

which represents the square root of the sum of the squared amplitude differences between the samples of each signal. Although the distance measure is simple to compute, it is highly susceptible to typical distortions on time series. When using ED, these distortions must be removed, otherwise, other methods, invariant to these distortions, should be used. Examples of distortions are amplitude and offset distortion, phase

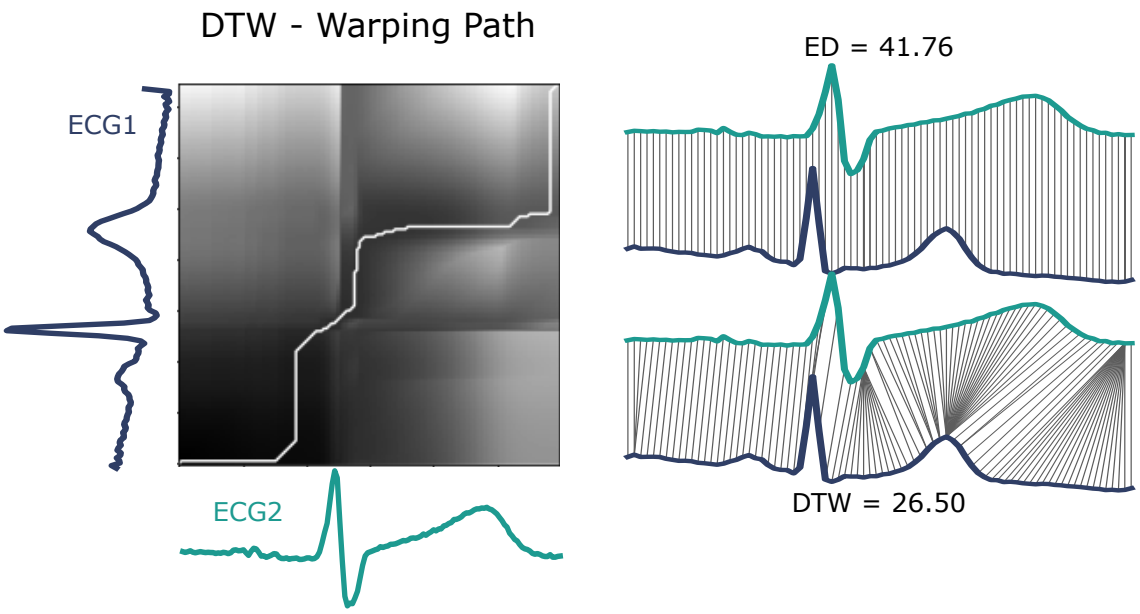


Figure 2.5: DTW and ED distances on two different ECG signals.

distortion, and local scaling ("warping") distortion. The first can be compensated by the z-normalized ED [17]:

$$z\_ED(Q,C) = \sqrt{2m(1 - \frac{\Sigma_{i=1}^{m} Q_i C_i - m\mu_Q\mu_C}{m\sigma_Q\sigma_C})}, \tag{2.8}$$

where μ_Q and μ_C are the mean of the time series pair and σ_Q and σ_C are the standard deviation.

The *warping* distortion can be solved with an elastic measure. For this purpose, DTW is typically used.

2.5.2 Dynamic Time Warping

The DTW distance measures the alignment between two time series. Let us consider two time series, Q and C, of length n and m, respectively:

$Q = q_1, q_2, ..., q_i, ..., q_n$
$C = c_1, c_2, ..., c_j, ..., c_m$

The alignment is measured by means of a distance matrix with size n-by-m, where the (i^{th}, j^{th}) cell of the matrix contains the $d(q_i, c_j)$ between the two points q_i and c_j, being $d = (q_i - c_j)^2$ [91]. Figure 2.5 shows an example of a distance matrix between two time series. The matrix fully describes the difference between the two time series and maps where these align. The mapping is made by a warping path, W, that represents the set of matrix cells that minimize the warping cost, also defined as the cumulative distance of these cells [91].

$$W = w_1, w_2, ..., w_k, ..., w_K; \qquad max(m,n) \leq K < m + n + 1, \tag{2.9}$$

$$DTW(Q,C) = min\sqrt{\Sigma_{k=1}^{K} w_k}. \tag{2.10}$$

The cumulative distance $\gamma(i,j)$ is calculated as $d(q_i, c_j)$ of the current cell added to the minimum distance adjacent to that cell:

$$\gamma(i,j) = d(q_i, c_j) + min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}. \tag{2.11}$$

When two time series with the same length have a linear warping path, such that $w_k = (i,j)_k, i = j = k$, we have a special case of the ED. DTW has a time and space complexity of $O(nm)$ while the ED has linear complexity ($O(n)$). Figure 2.5 shows an example of applying the ED and DTW on two different *PQRS* complexes from different ECGs.

2.5.3 Complexity Invariant Distance

A different type of distance measure is also used to cope with complexity invariances. This distance uses a complexity correction factor (CF) with an existing distance measure, such as ED [17]:

$$CD(Q,C) = ED(Q,C) \times CF(Q,C). \tag{2.12}$$

The CF is defined as [17]:

$$CF = \frac{max\{CE(Q), CE(C)\}}{min\{CE(Q), CE(C)\}}, \tag{2.13}$$

where CE represents the complexity estimate of a time series. This estimate is calculated based on the intuition that if we could "stretch" a time series until it becomes a straight line, this line would be as long as the complexity of the signal. It can be computed as the sum of the $n-th$ discrete differences along the time series [17]:

$$CE(Q) = \sqrt{\Sigma_{i=1}^{n-1}(q_i - q_{i+1})^2}. \tag{2.14}$$

These distance measures are performed on the original representation domain of the time series. As we showed above, other representation techniques can be employed, creating opportunities for other types of approaches. In this work, we propose other representation techniques to create novel ways of exploring time series. Therefore, other distance measures that can be used in different representations of time series will be explained.

2.5.4 Feature-based Distance

As mentioned, a feature series F can be computed from the original time series to represent it based on a specific feature. Each *subsequence* scanned by the *moving window* is characterized by the feature value, and a F is computed as an array. When multiple features are extracted, each *subsequence* is characterized by a set of features, creating a feature vector $\vec{f}$ with r feature values (to be clear, a feature vector is the set of feature values for a *subsequence* of the time series, while a feature series is a feature representation of the entire time series).

Vector-based distance measures can be used with feature vectors to compare different time series or *subsequences*. There are several vector-based distance measures, including the already mentioned ED or the manhattan distance, but we will only describe the cosine similarity/distance.

The cosine similarity is a measure of the angle between two vectors determining if these are pointing in the same direction. Consider two feature vectors $\vec{f_A}$ and $\vec{f_B}$. Their cosine similarity is computed as their normalized dot product [65] (equation 2.15).

$$CS = \frac{\vec{f_A} \cdot \vec{f_B}}{||\vec{f_A}||||\vec{f_B}||} \tag{2.15}$$

being $||\vec{f_A}||$ and $||\vec{f_B}||$ the euclidean norm of each feature vector, defined as $\sqrt{\Sigma_{i=1}^{r} f_{Ai}^2}$ and $\sqrt{\Sigma_{i=1}^{r} f_{Bi}^2}$, respectively [65]. Because the cosine similarity includes a vector normalization, it does not rely on the magnitude of the vectors and the difference is solely due to the angle between vectors. This is relevant for domains where vectors have high dimensionality, which is the case for feature vectors.

2.6 Applying Distance Measures

Measuring distances between time series allows to compare them. It is the fundamental instrument for most time series data mining tasks. With a distance measure, we can compare groups of time series for classification purposes or compare *subsequences* with a query template to find if it occurs in the time series. Not only can we compare different time series, but also compare the time series with subsequences of itself. By comparing each of its *subsequences* to all other *subsequences*, relevant structural information becomes available. In this subsection, relevant methods applied with the help of the presented distance measures are explained to retrieve information from a time series. We will start with distance/similarity profiles.

2.6.1 Distance Profile

As mentioned in the previous subsection, measuring all the distance pairs of a time series provides the ability to retrieve relevant structural information. When computing the

distance of a *subsequence* to all the other *subsequences* of the time series, a *distance profile* is calculated. Each *subsequence* can have a *distance profile* and when computing all the distance pairs, a self-distance matrix is the result.

Recently, a strategy was proposed to compute a one-dimensional distance *profile* for a time series based on a z-normalized ED matrix. By keeping the lowest value of each *distance profile* (nearest *subsequence*), we retrieve the *matrix profile* [58]. The result gives the minimal distance pair of each *subsequence*, meaning that minimum values are *motifs* and maximum values are *discords*.

Motif: The *subsequence* pair that has the lowest distance forms a motif. That means that on the entire time series, these two *subsequences* are the closest ones. The opposite is a *discord*.

Discord: The *subsequence* pair that has the highest distance forms a *discord*. That means that on the entire time series, these two *subsequences* are the furthest apart.

2.6.2 Self-Distance Matrices

A time series can reveal relevant information when each *subsequence* is compared to all the other *subsequences* of the same time series. The result is a pairwise distance matrix that unveils *homogeneity*, *repetition* and *novelty* on the time series [133]. Each are relevant for segmentation and summarization tasks.

Let X be a sequence with size $N \in \mathbb{N}$ that can be a time series or a representation of a time series in the PAA or feature space, such that $X = (x_1, x_2, ..., x_i, ..., x_j, ..., x_N)$. Each element of X can either be a single value or a vector with r features. Independently of that, a matrix SDM with size $N \times N$ can be computed, such that:

$$SDM(i, j) = d(x_i, x_j), \tag{2.16}$$

being d a distance measure between elements $x_i, x_j \in X$ for $i, j \in [1 : N]$. $SDM(i, j)$ represents a cell of SDM that contains the distance value. When $i = j$ the distance should be zero, therefore the diagonal of SDM has the lower values. Besides the main diagonal, other relevant structures can be found in SDM. These include *homogeneous blocks* and *paths* [133, 134], as can be seen in Figure 2.6.

Areas with lower distance and block structure along the diagonal are highlighted as *homogeneous* segments. These give an indication of *homogeneity* and *novelty*. *Homogeneity* because a *block* along the diagonal means that the time series has a constant behaviour during the segment delimited by the *block*. *Novelty* because when SDM has multiple *blocks* along the diagonal, it shows that the time series shifted its behaviour/regime. The moment there is a transition between *blocks* is a potential segmentation point. In Figure

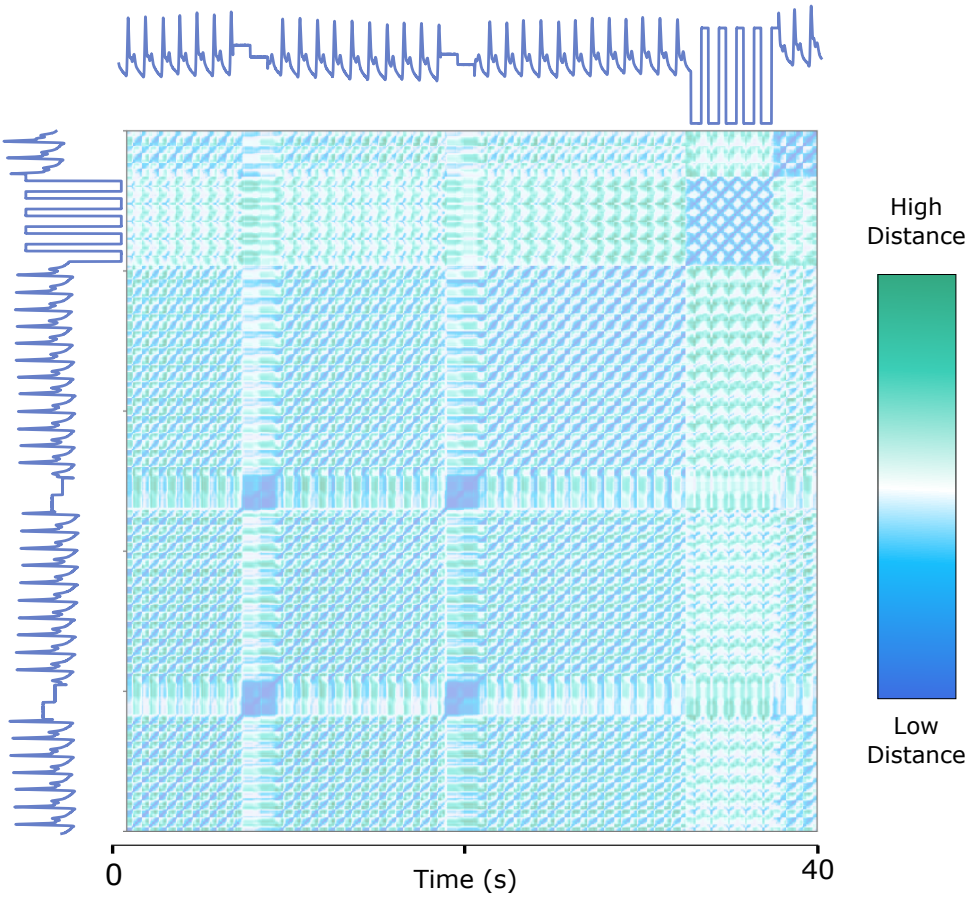


Figure 2.6: Self-Distance Matrix (SDM) representation of the ABP signal. The matrix was computed using the cosine distance between the feature representation of the signal. The main structures that can be highlighted are *blocks* along the main diagonal of the matrix, *paths* that are parallel to the main diagonal, and *cross-paths* that are perpendicular to the main diagonal. The signal comes from a public dataset (see Section A.1.2).

2.6, changes in the ABP signal are visible as blocks. In total, seven regimes are present on the signal, and seven blocks can be counted along the main diagonal.

Repeating *subsequences* in the time series are represented by *paths* on SDM. *Paths* are observed on the SDM because of matching *subsequences* during the sliding process, which computes consecutive low distance values represented in a diagonal form on the matrix. If the *subsequences* being compared are exactly the same, the diagonal is perfect, but differences might be found, which results in a distorted diagonal. In Figure 2.7 we show an example of an SDM with a chirp signal that has a continuously increasing frequency. A *subsequence* of the signal is compared with itself (Figure 2.7.left) and with the entire chirp signal (Figure 2.7.right). We highlighted the relevant *paths* in orange. When comparing the signal with itself only, the *path* is a perfect straight line, but when compared with the entire signal, the *paths* are continuously more distorted and shorter. These show where each period of the sine wave occurs, and the distortion results from the increasing frequency of the signal. In the SDM from Figure 2.6, *paths* are visible on the matrix, indicating the

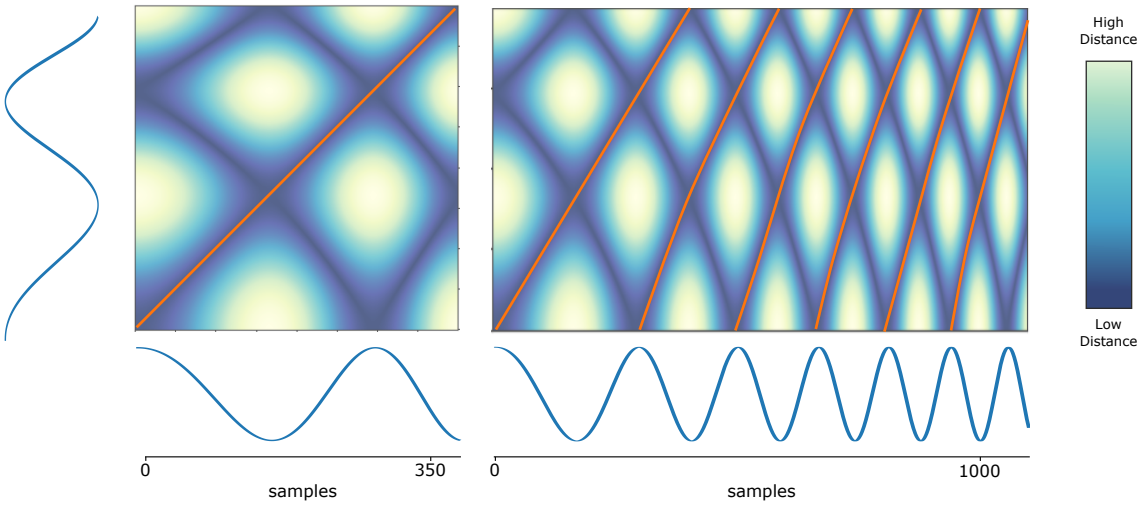


Figure 2.7: SDM of a chirp signal computed with the pairwise euclidean distance. The chirp signal is a sine wave with increasing frequency (in this case from f_0=1Hz and f_1=25Hz)[191]. Left) *Subsequence* of the chirp signal compared with itself. Right) The same *subsequence* is compared with the entire chirp signal.

presence of periodicity. The *cross-paths* indicate the same periodic behaviour, but also that the *subsequence* is symmetric.

It is relevant to highlight that as distance matrices can be computed, similarity matrices can as well, by following the same equation 2.16, but using a similarity measure (s) instead of d. Further in this manuscript, we will only mention the similarity matrix, which will be called SSM, as the proposed method uses the cosine similarity.

2.6.3 Template-based Search

The presented distances can also be used to retrieve a distance profile from a *template*. This type of mechanism belongs to the class of query-based search problems. The process to compute this distance profile involves sliding the template along the signal and applying a distance measure to each iteration. The result should indicate which *subsequences* are more (dis)similar to the used template.

An example of using a query-based search is illustrated in Figure 2.8, where a *PQRS* complex of an ECG is used as a template to search for all the other complexes. The distance computed is the z-normalized ED. The result shows a distance profile from which *minima* indicates the match with the template.

Other methods can be used to perform query-based search tasks, even using other types of templates, which can be a *drawing* [123] or text [4]. In this work, we will introduce novel ways of performing a query-based search with regex and natural language.

2.6.4 The k-Nearest Neighbors

Having distance measures we can compare signals for several purposes, namely classification. One traditional supervised algorithm for this purpose is the k-Nearest Neighbor

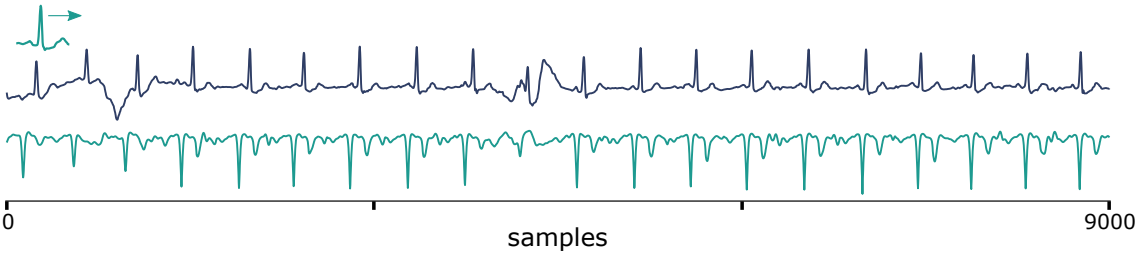


Figure 2.8: Distance profile of a query based search using the z-normalized ED. The template was a PQRS complex of an ECG.

(NNbr). The assumption of this method is fairly straightforward: new examples will be classified based on the class of their NNbr, that is, the class of the example is the average of the class of its k-NNbr. The process has two steps: (1) finding the k-NNbr and (2) choosing the class of the example based on the neighbors [41].

In order to find the NNbr, the distance from the example to all the time series of the training set has to be computed. The k-NNbr are the k training examples with the lowest distance. Having the NNbr, the next stage is the determination of the example's class, which can be made with several strategies, such as majority voting, or distance-weighted voting [41].

In this work, we will propose a novel method for time series classification that will have its performance compared with a 1-NNbr based on the z-normalized ED. The proposed method is from the text-domain. As we have been mentioning in this manuscript, parts of the work developed are tightly connected with text mining methods by making a bridge between time series and text. It is then relevant to explain this relationship and how we can use it for time series data mining applications.

2.7 Text Mining on Time Series

2.7.1 Time Series Textual Abstraction

In SAX (see Section 2.4.4), the signal is transformed into a sequence of symbols. For this, each sample of the PAA representation is converted into a *character*, which can then form *words* and *sentences*. As a novel symbolic representation of time series is proposed in this work, it is relevant to give the general background that helps make the association between the original time series, how it can be transformed into a symbolic time series, and the text notation that can be used for this purpose. Note that this is an introductory explanation that will be further contextualized when needed throughout this manuscript. We start with fundamental definitions.

Character: A *character*, c, is an unit symbolic element belonging to the vocabulary V that represents a sample or *subsequence* of a time series, such that $c \in V$. Each sample of a time series is transformed into a *character* to form a *symbolic time series*.

Symbolic Time Series: A *symbolic time series* is a sequence of *characters* ordered in time generated from the parent time series. It has length $n \in \mathbb{N}$: $ST = (st_1, st_2, ..., st_n)$, with $st_i = c \in V$. A specific sequence of *characters* of a *symbolic time series* can form a *word*.

Word: A *word* can be a single character or the concatenation of a sequence of *characters*, giving a textual representation of a *subsequence*, such that $W = (w_1, w_2, ..., w_u)$, with $u \in \mathbb{N}$, $w_i = c \in V$ and $W \in V$. Putting *words* together forms a *sentence*.

Sentence: A *sentence* S represents a group of *subsequences* of the time series, or the entire time series itself. It is formed by joining sequences of *words*, such that $S = W_1, W_2, ..., W_s)$, with $s \in \mathbb{N}$ and $W_i \in V$.

Document: The set of *sentences* in a time series are called a *document*. The *document* D can be computed by one or more sentences. In this work, a *document* represents the textual description of the entire time series.

Corpus: The *corpus* is a collection of text material (group of *documents*). It represents a higher level of textual information that is used in case we have a dataset with multiple time series. This collection is typically annotated and used for machine learning tasks. In this case, a corpus will be represented by the set of *documents* that describe an entire time series dataset.

Vocabulary: The *vocabulary* V comprehends the set of all different *characters* and *words* present in all time series of the dataset.

2.7.2 Text Features

Here are introduced traditional methods applied for feature extraction of text data, namely the BoW and TF-idf.

Bag of Words: A BoW is a feature matrix representation of a corpus, being the feature the number of occurrences of each *term* (in this case it can be a *character* or a *word*), called the tf:

$$bow(t, d) = tf_{t,d} = \frac{f_{t,d}}{\sum\limits_{t' \in d} f_{t',d}}, \tag{2.17}$$

being t the term that exists in a document, d the document and t' the term that belongs to document d. Here t can be a single *word* or an *n-gram*.

N-gram: It is a span of followed *words* that are counted in the BoW/TF-idf. As an

example, a possible *2-gram* from Figure 2.4 would be `aa`, `ca` or `ac`. This strategy resembles a *moving window* with total overlap on time series, but for text. It can make the BoW/TF-idf model more robust since it relies in more than single *word* statistics. It also has a very good fit for time series, considering that it can take into account time dependencies between *words*, which reflects the time dependency seen in time series between *subsequences* (e.g. it might be more meaningful to say that a signal has a `peak` next to a `valley` than just saying that it has a `peak` and a `valley`).

The BoW is commonly used to vectorize the textual representation of each symbolic time series, but there is common knowledge in the text mining community that if a *term* occurs in all *documents*, then it is less relevant. In order to counteract this possible limitation, the TF-idf matrix is used.

Term Frequency Inverse Document Frequency: The TF-idf matrix increases the relevance of t by means of the t_f, while reducing its importance in proportion to the number of *documents*, d that contain the term t. The model is defined by being a ratio between the t_f and the *inverse document frequency* (idf), which is calculated as follows:

$$idf(t, D) = \log \frac{L}{|\{d \in D : t \in d|\}} \tag{2.18}$$

L is the total number of documents ($L = |D|$). The final equation of the *tfidf* model is the following:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{2.19}$$

Both BoW and TF-idf are matrices that have a vector representation of each *document*, where each element of the vector is the relevance of the *term*. That means that the euclidean or cosine distance (among others) can be used to compute the difference between *documents*.

Due to the probabilistic nature of the BoW and the fact that it contains discrete features, it is suitable to use in naïve Bayes classifiers. In the other end, the TF-idf is typically used with linear SVM classifiers [150].

2.7.3 Text-based Classifiers

As mentioned in the previous Section, linear classifiers, such as naïve Bayes and SVMs, are conventionally used for text document classifications. We will briefly explain the mechanisms of each of these classifiers that will be used in this work.

2.7.3.1 Naïve Bayes Classifier

Naïve Bayes classifiers are a set of probabilistic linear classifiers that follow Bayes' theorem [197, 154], which implies that the probability of an event occurring is based on *prior*

knowledge of conditions that might be related with this event. The term *naïve* is used because these classifiers assume that features are mutually independent.

The probability of an event translates as the probability of an object belonging to a specific class (w_j) given the observed features (x_i - feature vector). This probability ($P(w_j|x_i)$ - posterior probability) is given by the general equation based from Bayes' theorem [197, 198, 154]:

$$P(w_j|x_i) = \frac{P(x_i|w_j)\dot{P}(w_j)}{P(x_i)}. \tag{2.20}$$

- $P(x|w_j)$ is the class-conditional probability (likelihood), which calculates *how likely* the feature is observed given that it belongs to class w_j. It includes the probabilistic contributions of all feature dimensions:

$$P(x|w_j) = P(x_1|w_j)\dot{P}(x_2|w_j)...\dot{P}(x_n|w_j) = \prod_{k=1}^{n} P(x_k|w_j) \tag{2.21}$$

- $P(w_j)$ is the *prior probability*. The general probability of a class. In case the classes are uniformly distributed, the posterior probability is dependent on the class-conditional probabilities.

- $P(x)$ is the *evidence*, which is the probability of encountering feature x independently of the class label.

During the training process, the posterior probability is maximized. Several approaches can be found to compute naïve Bayes models, namely *Gaussian*, *Multinomial*, *Complement*, *Bernoulli* and *Categorical* naïve Bayes. In this manuscript, we will show the usage of a *Multinomial* naïve Bayes model for classification, which uses word count feature vectors and is more appropriate to be used with BoW or TF-idf models. We will use the *Python* implementation given by the *scikit-learn* package called *MultinomialNB* [150].

2.7.3.2 Linear Support Vector Machines (SVM)

The SVM classifier is a conventional method that separates different classes by finding the *hyperplane* from n-dimensional space (n equal to the number of features) that puts all samples from the same class the furthest apart from the other classes' samples. The training process has the purpose of finding this hyperplane by maximizing the margin distances between the hyperplane's surface and all samples. The hyperplanes can be written as the set of features x that satisfy the following equation:

$$\mathbf{w}^\top \mathbf{x} - b = 0, \tag{2.22}$$

where $\mathbf{w}$ are the weights for the optimal hyperplane, given as a linear combination of support vectors [37]. The support vectors belong to samples that give shape and orientation

to the hyperplane. There are several ways of performing the maximum-margin hyperplane optimization, originally linear, but with non-linear metrics also available. In this work, we used the linear SVM implementation available at the *Python* implementation given by the *scikit-learn* package called *LinearSVC* [150].

2.7.4 Text Pattern Search

When writing or reading a document, we often use shortcuts or other commands to speed up the task of searching for specific words or expressions. In some cases, we might not be searching specifically for a character or word, but rather for a text structure or text pattern (for example, one might be searching for all the emails available in a text document). This type of search is comparable to querying in time series, being the most convenient method to perform this type of search in text a regex.

A regex is a parsing technique that is convenient to write text patterns, being more flexible than direct matches. It is based on regular languages, follows a specific set of rules, and contains a set of meta-characters.

In order to understand the way that regex work, some of the most used characters and regex primitives are presented as follows [52]:

2.8 Performance and Validation Measures

This manuscript discusses three main time series data mining problems, namely classification, segmentation, and event/pattern detection. In order to perform a validation of the work developed, standard procedures are used. In this section are explained which procedures are typically employed to evaluate the performance of algorithms in the mentioned problems.

2.8.1 Classification Problems

One of the most common strategies to evaluate algorithms from the machine learning field are *precision* (P), *recall* (R) and *f1-score* (F1). These measures are calculated based on *true positives* (TP), *false positives* (FP) and *false negatives* (FN). Understanding what these metrics represent depends on the problem. These measures are used in this thesis to evaluate the performance of classification and event detection algorithms.

In time series classification problems, a labeled dataset is divided into training and testing sets. The algorithm learns on the training set and a final evaluation is made on the testing set. In case the algorithm needs to fine-tune *hyperparameters*, it is good practice to perform a cross-validation during the training stage. This ensures that the algorithm is optimized without *seeing* any testing samples and prevents *overfitting*. Several strategies can be used to perform the separation of the training set, such as the *k-fold* or *leave one out*. Besides, these can be made by shuffling, stratifying or grouping it. Choosing the most adequate cross-validation method depends on the distribution of the training samples.

MC	Description	Example of Match
*	The preceding item will be matched zero or more times	*eve*nt* → [evnt, event, eveent]
+	The preceding item will be matched one or more times	*su+b* → [sub, suub, suuub]
?	The preceding item is optional and will be matched, at most, once	*team?* → [tea, team]
.	Matches any character	*s.m* → [ssm, sam, sim]
[]	Matches anything inside the brackets	*wom[ae]n* → [women, woman]
|, &	Boolean operators - or, and	*tr(i|a)p* → [trip, trap]
(?=<)	Positive lookbehind - The string matches the item that is preceded by the pattern inside the lookbehind without making it part of the match	*(?=<http://)* → any URL
(?<!)	Negative lookbehind - The string matches the item that is not preceded by the pattern inside the lookbehind	*?<!\ d*.+* → [~~10th~~, th]
(?=)	Positive lookahead - The string matches the preceding item that is followed by the pattern inside the lookahead without making it part of the match	*(?=www\ .)* → matches web protocol
(?!)	Negative lookahead - The string matches the preceding item that is not followed by the pattern inside the lookahead	a(?!b) → [~~ab~~, ac]

Table 2.1: Main regex operators and meta-characters. Each is presented with a simple example of a good match for a possible regex. A description is also made for complementary understanding.

In order to perform the validation or final evaluation of the algorithm, the ground truth labels are compared with the labels predicted by the algorithm on the testing set. This comparison gives the number of *TP*, *TN*, *FP* and *FN*.

- TP_c - If the label predicted by the algorithm is equal to the target class (positive when positive);
- TN_c - If the label predicted is correctly not the target class (negative when negative).
- FP_c - If the label predicted by the algorithm is falsely classified as the target class when it should not (positive when negative);
- FN_c - If the label predicted by the algorithm is not labeled as the target class when it should (negative when positive).

$$Precision(P) = \frac{TP}{TP + FP} \tag{2.23}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{2.24}$$

$$F1 - score(F1) = 2 \times \frac{P * R}{P + R} \tag{2.25}$$

$$Accuracy(A) = \frac{TN + TP}{TN + TP + FP + FN} \tag{2.26}$$

These measures were initially performed in binary classification problems, but can be adapted in multi-class ones as well. For this, each class (the target class) is compared to all the other classes, being the target class the *positive* and all the other classes *negative*. All measures are calculated for each class, and a macro and micro average of these metrics are finally calculated (here c represents the number of classes).

$$macroP = \Sigma_{i=0}^{c} \frac{P_i}{c} \tag{2.27}$$

$$macroR = \Sigma_{i=0}^{c} \frac{R_i}{c} \tag{2.28}$$

$$microP = \frac{\Sigma_{i=0}^{c} TP_i}{\Sigma_{i=0}^{c} TP_i + \Sigma_{i=0}^{c} FP_i} \tag{2.29}$$

$$microR = \frac{\Sigma_{i=0}^{c} TP_i}{\Sigma_{i=0}^{c} TP_i + \Sigma_{i=0}^{c} FN_i} \tag{2.30}$$

These metrics are typically supported with a *confusion matrix*, which displays a 2D-map of the ground truth labels *versus* predicted labels.

2.8.2 Event Detection

Regarding event detection problems, the process involves finding the sample that corresponds to the ground truth event. Considering that it would not be fair to calculate the performance of an event detection algorithm only by searching if the ground truth sample exactly matches the ground truth, we calculate the TP, FP and FN based on a tolerance margin. From these measures, metrics from Equations 2.23, 2.24 and 2.25 are calculated. The estimated events are considered one of the following categories:

- TP_e - is counted when the estimated event is in the margin around the ground-truth event;
- FP_e - is counted whenever it is out of a margin around the ground-truth event, or when there is more than one estimated event inside the margin;

- FN_e - is counted when there is no estimated event inside the margin of the ground-truth events.

Additionally, the distance of the TP events from the ground-truth events can be calculated with the Mean Absolute Error (MAE):

$$MAE = \sum_{i=1}^{k} \frac{|g_i - e_i|}{k} \tag{2.31}$$

The precision measure is relevant to indicate if the method can only estimate events that belong to the ground-truth category, while the recall measure is an important indication of how many ground-truth events are missed in the estimation of the method. Both measures are combined in the F1-measure.

2.8.3 Evaluating Query Complexity

As we are introducing novel ways of performing more expressive query-based searches with regex, we are measuring the legibility and difficulty in generating a query. In the programming field, *Halstead* measures are typically used for this purpose. These measures describe the complexity of the script directly from the source code, based on a set of metrics calculated with the number of distinct operators (oprt) and operands (oprd), and the total number of operators (Toprt) and total number of operands (Toprd). These metrics are the [89]:

Vocabulary The number of distinct operators and operands that belong to the script:

$$Voc = oprt + oprd \tag{2.32}$$

Length

The total number of operators and operands that belong to the script:

$$Lgth = Toprt + Toprd \tag{2.33}$$

Calculated Length

It uses the entropy measure to calculate the average amount of information based on the number of distinct operators and operands:

$$CL = oprt * \log_2(oprt) + oprd * \log_2(oprd) \tag{2.34}$$

Volume

It measures the amount of information that the reader has to absorb to understand its meaning. It is proportional to the length measure ($Lgth$) and logarithmically increases with the vocabulary:

$$Vol = Lgth * \log_2(Voc) \tag{2.35}$$

Difficulty

The difficulty in writing or reading the script. It increases more having fewer operands repeated more frequently than having more operands repeated more frequently:

$$Dif = (\frac{oprt}{2} + \frac{Toprd}{oprd}) \tag{2.36}$$

Effort

Measure of the effort necessary to understand what is written and recreate the script. It is proportional to both volume and difficulty measures:

$$Efrt = Vol * Dif \tag{2.37}$$

2.8.4 Comparing Algorithms' Performances

In science, a good practice is to compare the proposed algorithm with other existing solutions, such that the reader can understand how the results presented are unbiased from the data. In this work, each strategy proposed in the domains of classification and event detection will be compared with existing solutions. Typically, the results will have accuracy or F1-scores metrics, sometimes from very large datasets. Averaging these scores may not give a reliable view of performance differences between methods, specially if these scores are the performance of the algorithm over different datasets, with different sizes, scenarios and difficulties. In that case, other techniques should be used to present the results such that the reader can have the full picture.

One possible strategy involves counting the times a method wins, draws or loses against another method. This can be helpful to have a broader picture of the results in addition to the average metrics. We will use this count strategy for both classification and segmentation tasks since we will be using multiple datasets, with different lengths, difficulties and contexts.

Visual tools can also be employed for this purpose, namely *confrontation maps* and *critical distance maps*, with examples displayed on Figures 2.9.left and 2.9.right, respectively.

2.8.4.1 Confrontation Maps

When the proposed algorithm is applied to a dataset and compared with another algorithm, we might be tempted to display an overwhelming quantity of information in a table. Although this is a valid approach that should be made to give a full picture to the reader, there should also be a straightforward way of displaying the same information, but reading it at a glance. With confrontation maps, we can compare the performance of two algorithms just to very intuitively understand if there is a significant difference in their performance. Typically, this map is a scatter plot comparing the *F1-score* or *accuracy* of algorithm 1 *versus* algorithm 2. Figure 2.9.left displays an example of it taken from [90]. Each dot of the plot is a dataset. When it is above the diagonal, it is better classified by

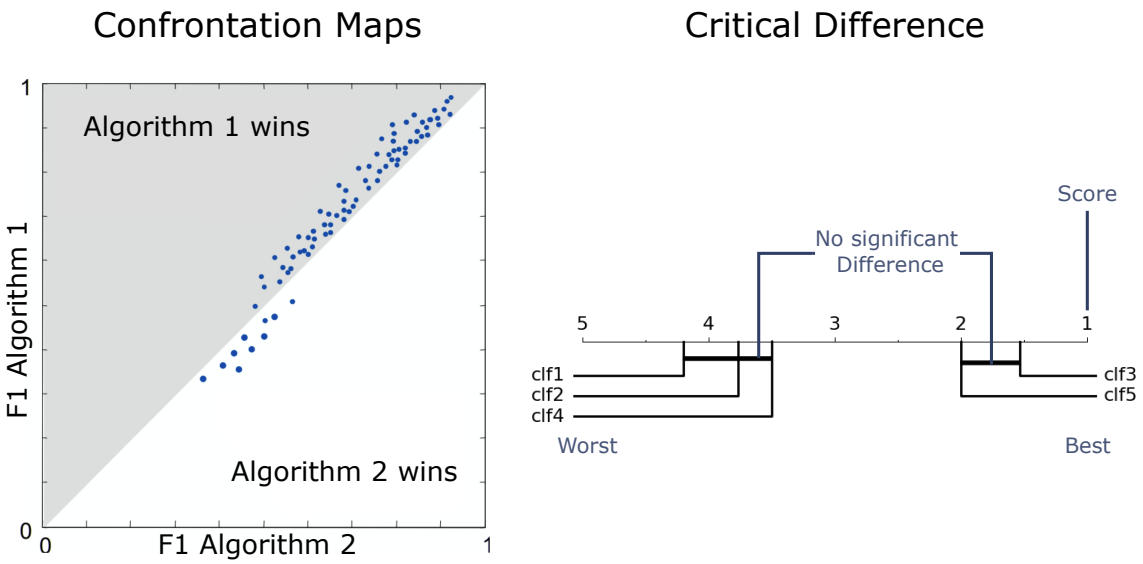


Figure 2.9: On the left are confrontation maps, used to compare classification algorithm performances. On the right is a critical difference map, which shows a statistical difference in the performance of algorithms for general purposes.

algorithm 1, while if below, it is better classified by algorithm 2. In this example, algorithm 1 is better in most datasets.

2.8.4.2 Critical Distance

Critical distance maps are a way of comparing the performance of multiple algorithms or variations of the same algorithm. The plot is a representation of a statistical test over the performance result of each algorithm. The test evaluates if the difference in the performance is significant (critical distance) or not. For instance, in Figure 2.9.right, the plot compares five different classifiers (*clf1* to *clf5*) and highlights that the difference in performance is not significant between *clf1, 2* and *4*, neither between *clf3 and 5*. However, *clf3* and *5* have a much better performance than the other classifiers. The closer the classifiers are from the right (1), the better they are. The tick bar connects the classifier with performances that are not significantly different in the statistical test.

This method receives the performance scores (such as F1-scores, accuracies, etc...) of multiple classifiers. This type of comparison gives a more appropriate explanation of the differences between different methods than a simple average. This is even more relevant in cases where the methods are applied in multiple datasets that have different size, lengths, and/or difficulty, and where the average performance results would be not fair to describe the differences.

In this work, we will use an implemented critical distance method from [84], which uses the *Wilcoxon-Holm* test to counteracts the problem of multiple comparisons and calculates pairwise significance between all methods evaluated [6]. The null hypothesis of the *Wilcoxon* signed rank test is that the median of the population of differences between the paired data is zero, while the alternative hypothesis is that it is not. The significance is

then evaluated with the *Holm* test to accept or reject the multiple hypothesis results based on the p-values resulting from the pairwise tests.

3

State of the Art

In this Chapter, an extensive delivery of the existing works regarding the main topics covered in this thesis are presented. This will include (1) existing works for information retrieval in time series, mainly in terms of event detection and segmentation, (2) approaches for summarization of time series, (3) symbolic representation techniques and how these can be used for (5) classification tasks and (6) query-based search mechanisms.

3.1 Novelty, Periodicity and Similarity Search on Time Series

Prior works in information retrieval, most specifically in novelty, periodicity and similarity search, focus on change point detection or segmentation, where the strategies are categorized as online versus offline, univariate versus multivariate, model-based versus non-parametric and unsupervised versus supervised [26, 8, 187]. We will further explore the existing literature in supervised and unsupervised methods for this purpose. In addition, as the proposed approaches rely on the usage of a SSM, the currently available knowledge on the SSM for the same tasks are explained.

3.1.1 Supervised Methods

Supervised methods include multi-class, binary and virtual classifiers optimized to detect change points [8], where the nature of the change can be provided as an additional advantage. Emerging works are found to focus specifically on biosignal segmentation, e.g. applying neural networks (NNet) to ECG signals. In *Pedro Matias et. al*, an NNet with transfer learning was used for the segmentation of periodic biosignals (motion and ECG) [125]. Convolutional NNet has also been found for ECG segmentation. In *Pedro Silva et. al*, a convolutional NNet on a binary classification task (*heartbeat* or *not a heartbeat*) was proposed [179], while *Aman Malali et. al* put forward a convolutional long short-term memory (LSTM) NNet for the same task [120]. Last but not least, *Viktor Moskalenko et. al* used a UNet-like full convolutional NNet for the ECG signal segmentation of P and T waves, as well as the QRS complex [132]. Without model training, the ECG segmentation

can also be solved through a subsequence search in the context of a carefully selected query pattern [48].

Wearable technology has also improved the field of gait analysis, increasing the interest in gait event detection [53]. Recently, machine learning approaches have been found for gait segmentation related to Parkinson's disease [54], and hidden Markov models (HMM) were also used for the same purposes [165]. Gait event detection methods based on rapid positive changes in GYR data were employed for rehabilitation research [124, 28]. The work from *Matteo et. al* shows the ability of deep learning techniques to improve gait segmentation [53]. Traditional signal processing methods, such as the integral of the signal envelope, can be applied to EMG signal segmentation for gait analysis [192].

Biosignals' segmentation also facilitates medical research. For the study of sleep staging segmentation, *Mathias Perslev et. al* introduced a fully convolutional network [151]. A square-root velocity function to segment periodic data for posterior alignment and statistical analysis helped disease classification based on the *Karcher* mean [97].

A problem reported by [8] states that supervised methods rely upon brittle training sets and class imbalance, since there are more in-state sequences than change point sequences [8]. An additional problem reported by [26] is that most algorithms' performance was validated in synthetic data, where the given nature of the application was not optimal. In response, a benchmark is available for change point detection [26], where methods can be compared on real data. This study applies this benchmark as a reference of state-of-the-art methods.

3.1.2 Unsupervised Methods

Existing classic non-supervised methods in change point detection, such as the *Bayesian Online Change Point Detection* (BOCPD) [**bocpd2**, 2], *Binary Segmentation* (BINSEG) [15], and *Segmentation Neighbourhood* (SegNeigh) [13], are witnessed to be able to perform state-of-the-art applications in various domains [26]. BOCPD is a non-supervised model-based method for change point detection that was simultaneously introduced by *Fearnhead and Liu* [**bocpd2**] and *Adams and MacKay* [2]. The method infers a change point based on the fact that the model parameters before and after the change point are independent. It relies on learning a joint probability distribution since the time of the most recent change point (*run length)* by means of a recursive message-passing algorithm. The calculated recursive probabilities will be tested to evaluate if the *run length* will be zero. If so, a change point is detected; if not, the *run length* continues to increase [2, 26]. This method needs hyperparameter tuning for sound performance [26]. The BINSEG method is a greedy sequential algorithm, recursively partitioning the signal into smaller segments. The position where the signal is segmented is typically selected where the cost function is minimized. BINSEG has not been reported to cope with a multi-timescale change [26, 15]. The available repository [26] collecting the implementation of some offline methods [187] above lacks a visual output that can provide users with the location of the change points.

In this work, we used the benchmark dataset available in this repository to compare the performance of the proposed solution with the mentioned methods.

Window-based segmentation, typically relying on a sliding window that is divided into two smaller windows based on the comparison using a cost function, can apply to real data domains [187]. Another approach, called *Fast Low-cost Online Semantic Segmentation* (FLOSS) [**floss**, 57], searches regime changes based on the nearest neighbours of subsequences, which allows the similarity comparison between segments for the segmentation and summarization of long-term time series.

Several works that use unsupervised methods are found for the identification of cyclic information and anomalies. An automated algorithm of segmentation was able to separate complex and multidimensional data into smaller segments that can be described through harmonic models. This algorithm revealed to be significantly useful to identify cyclic movement without any *a priori* knowledge of the input data, using a combination of a recursive least squares segmentation algorithm, a model fitting of damped harmonics, and in the end, a clustering analysis to classify the events [114, 113]. The usage of features is of great relevance in unsupervised works, and methods are found to select adequate features for detection and classification tasks, such as in [118]. Another example is the usage of four-pass UKF (unscented Kalman filter) to produce an unified model with kinematic parameters. These may then be segmented by analyzing the parameter's zero crossing velocity and in the end uses a clustering algorithm to identify repetitive segments [193]. Other methods rely on a self-similarity approach, namely in the work of *Neuza et. al*[140], where cyclic information is segmented by searching for minimums, in the convolution of a segment of the signal with itself. The *Matrix Profile (MP)*, which is a method that compares all sub-sequences of a given time series with themselves through the ED and returns the minimum value for each distance profile. This method highlights the moments of the time series which are similar within themselves [209], being then used for a mutltitude of purposes in time series data mining [59, 115, 81, 119]. Additionally, autocorrelation revealed itself an useful tool, as the search over maximum values can infer the cyclic nature of the data [18]. Finally, for anomaly detection in industrial scenarios, *Varandas et. al* applies an unsupervised method based on the clustering of time series segments to detect improper movements [190].

3.1.3 Self-Similarity Matrix (SSM)

Similarity matrices have been used for the analysis of time series in the past. These matrices have several representations and denominations, being the most commonly known *recurrence plot*. These plots are a similarity matrix thresholded by a specific value that highlight only similar regions. The representation of these maps was used to analyze time series [207, 212] and search for reoccurring shapes, anomalies or symmetric behaviours [207]. It was also used for time series classification by inputting the recurrence plot into a neural network (Convolutional Neural Networks) [212, 94] or forecasting [11].

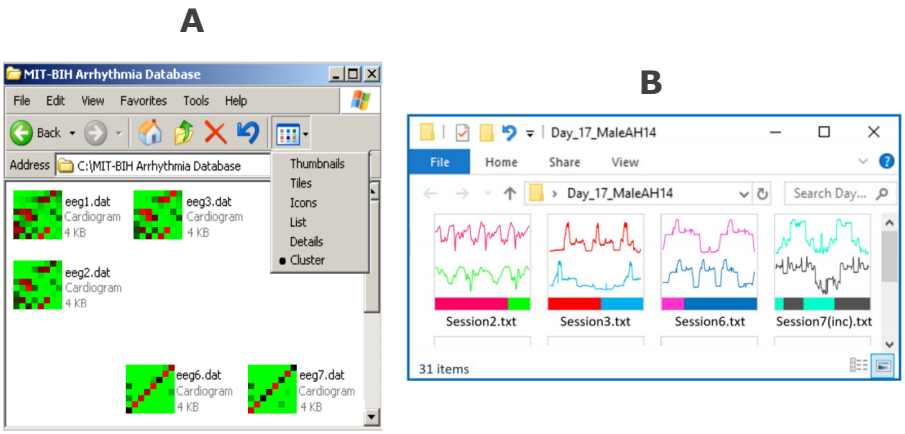


Figure 3.1: Strategies for time series summary found on the literature. These images are taken from the works from [82, 96]

The SSM has more information than the recurrence plot (which has been thresholded). Several usages have been explored in the audio domain, namely for segmentation, thumbnailing, periodiciy search or music alignment. For this, the audio signal is transformed on a feature-based representation [133, 134]. The advantage of the SSM is that it provides a considerable amount of information for a specific timescale. This study promotes SSM concepts and applications from the audio domain to other time series domains. The proposed method can detect events with context, associating the estimated events with patterns, (dis)similarities, periodicity and novelty, and a possible extension is the task of summarization.

3.1.4 Summarization

Very few strategies are found to make compact and meaningful representation of time series. The works that can be highlighted refer to time *snippets* and time series *bitmaps* [82, 96]. The first highlights the limitation of current methods in providing a satisfactory solution to time series *summarization*. It proposes a method that is able to segment the k most *representative* sub sequences of a time series, and use these elements as the summary. This strategy answers the segmentation and similarity. Regarding the time series *bitmap* representation, the strategy is able to provide a coded bitmap with information on cluster, anomaly and other regularities on data collection. These bitmaps were used as folder icons, and also answer several of the aforementioned characteristics, such as *similarity* and *events*. An example of both strategies can be seen on Figure 3.1.

Time series *shapelets* are also a method that could provide interesting results. However, the strategy is *supervised*, and the point of the proposed method is to have *no apriori* knowledge about the structure of the data, except the time scale in which the summarization is performed. Other interesting strategies provide a transformation of time series into text and could be used for time series summarization, but are not able to suitably summarize a time series from the textual representation [162, 103].

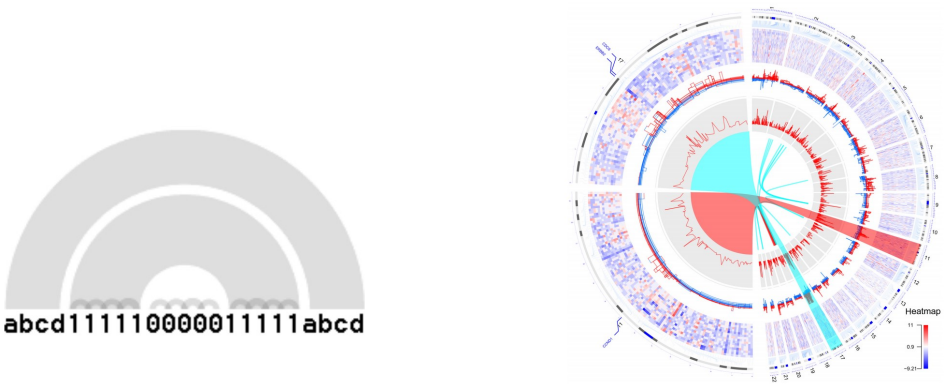
Strategies that are typically used to present information in a compact way are found in several domains. In text analysis, for instance, the relationship between repeating sequences is illustrated with arc diagrams [96, 196]. These show where repeating sequences occur in a very concise way. This has a range of applications that include, for example text and DNA sequence analysis.

One domain that has a particular relevance in data visualization is genomics. Graphical genome maps are found to concatenate a significant amount of information in a very compact way. Genome features and sequence characteristics are assessed with this visual strategy. An example can be found on Figure 3.2b. This visualization strategy can provide increasing circular layers of information. Although we are used to look at time series from left to right, a circular representation can have benefits to concatenate the information we want to include.

In the musical domain, strategies have also been developed that summarize audio time series with segmentation techniques. One of the strategies that is common to be used involves detecting novelty instances on a similarity matrix representation of the audio signal, called SSM. This data structure provides a significant range of information that can be used to retrieve structural information, such as block and periodic structures [133, 134, 147, 20]. This method will inspire our visualization strategy, which will be explained in a further section.

3.2 Text-based representation of time series

In the field of signal processing, few are the algorithms that rely on syntactic models to solve query search tasks. In the late 60s and 70s, syntactic approaches in time series started to appear, but have fallen into disuse over time and have recently re-emerged.



(a) (b)

Figure 3.2: A - Diagram for string association. This image is taken from the works from [196]; B - Circular plot by OmicCircos. Several layers (Circular tracks) identify genome position, expression heatmaps, correlation between expression and CNV, among other features. The image is taken from the works from *Ying Hu, et al.*. [77].

Some application of these linguistic models into mathematical domains have been led by T. Pavlidis [149, 148] and K. Fu [88], which have an exhaustive list of studies in pattern recognition and computer vision fields related with the symbolic description of 2D-shapes. Moreover, *Pavlidis* expressed that the primary problem in the implementation of such methods with functions of one variable would involve the representation of the waveform into a one dimensional string of primitives over a finite alphabet [148].

These concepts inspired multiple works that have emerged and applied a syntactic approach to pattern recognition tasks. Some examples can be addressed in the field of ECG analysis [73, 188, 186, 38]. *Horowitz et al..* describe that the detection of peaks in ECG signals can be achieved by specifying a set of primitives (one symbol or a group of symbols that represent one sample of the signal) able to be representative of the shape of the ECG peaks. The generated string is made by combining the information of the amplitude and the first derivative, which is thereafter parsed based on a set of context free grammar rules. This enables to detect the ECG peaks.

Similar approaches were made in *Skordalakis and Udupa et al..*, although in these cases, the set of primitives is coded with the length and slope of the line segment [188, 186]. The use of symbolic representations has also been found in biometric ECG by means of compression techniques [38].

Recently, approaches based on the symbolic representation of time series have reappeared and provided satisfactory and competitive results in comparison to other, more conventional, statistical methods. These results are an evidence that the use of small sets of primitives and grammatical rules are a possible way of describing large sets of patterns by efficiently representing the signal's structures into an ordered hierarchical sequence of characters. [64]. For instance, in the last decade, several works were found in the literature that used symbolic approximation of time-series. Some were found to model finite state automaton from the generation of symbolic time series, which were able to describe the probabilistic flow of data and with this detect patterns on time series[153, 152]. Both *Lin et.al.* with the SAX and *Senin et al..* used grammar based compress methods to detect anomalies by comparing subsequences of the signal [103, 175]. Another similar approach uses a symbolic representation of time series for information retrieval [100].

In [145], *Bag-of-words* models are used on a symbolic representation of time series to classify physiological data. Additionally, several works regarding symbolic time series analysis were also found for human gate dynamics analysis [1, 210] and EEG epileptic seizures and brain dynamics [78]. In [185] the symbolic representation of time series is used to uncover regions that share similar climate patterns by means of transition probabilities over the symbolic sequence. Finally, *Hamdi et.al.* used a novel approach that applies a deterministic finite automaton (DFA) and regular grammar rules for real time detection of the ECG's QRS complex [64].

Inspired by the stated syntactic approaches for pattern recognition and by the importance of having more interactive tools for time series analysis, we developed a methodology that comprises three major steps. These are in line with the ones typically found in the

bibliography: pre-process the time series to return a more appropriate signal; symbolic representation of time series into a set of primitives that give information about the shape and attributes of the time series over time and is governed by a set of grammatical rules; and a parser to search for patterns in the symbolic representation.

In this manuscript, we will propose a novel symbolic representation. Unlike the previously reviewed approaches, it aims to generalize the application of the syntactic apparatus to build a generic tool, so it can be easily tuned to solve problems that involve query search tasks applied to generic time series data.

3.3 Meaningful query-based search on time series

There is a large literature on time series similarity search, see [194] and the references therein. However, in most cases it is assumed that the query comes from a downstream algorithm, not a human. As such, there has been relatively little attention paid to the ability of humans to formulate meaningful queries. In principle one could do "query-by-sketching" and invite the user to draw the pattern he/she is interested in finding [123, 136]. The recent "Qetch" system is a prominent example of this approach [123]. However, classic distance measures may be too literal and limited in expressiveness to retrieve the desired pattern. As a simple example, suppose that a user wishes to retrieve all highly symmetric patterns. There is simply no way to do this with ED or similar distance measures. Other researchers have noted these issues and proposed more flexible query-based languages for time series. For example, the *SDL* (shape definition language) of [4]- allows the user to formulate "blurred" queries. However, we believe that most of such systems are not accessible for the typical user. For example, with SDL, expressing the search of three consecutive peaks would require: `(Shape triplespeak (width ht1 ht2 ht3) (in width (in order spike (ht1 ht2 ht3) spike (ht1 ht2 ht3))))`. In contrast, the solutions we will present further have a more flexible approach. There have also been a handful of other attempts at natural language querying for time series [79, 80], but none of these works seem to have been adopted by practitioners.

3.4 Classification of time series

Current available methods for time series classification are categorized as shape-based and structure-based. Existing approaches until the last decade were focused in shape-based similarity methods, while during the mid 2010's, methods that would seek the analysis of higher-level features started to be developed [92].

Shape-based methods focus their attention in performing local comparisons between time series. Examples of well-known methods are the ED or DTW [102]. Although both work well with short-length time series, the first has the inconvenient of needing time series with the same length, while also being sensitive to time misalignments. The latter is able to counteract this problem by means of determining the best alignment between

two time series [92, 102]. These distance measures are usually combined with a k-Nearest Neighbour (k-NN) classifier. The limitations of these techniques come with problems that include the presence of noise or long time series with characteristic sub-structures [169].

In the other end, structure-based methods rely on broader aspects of time series such as the presence of specific morphological structures or patterns, being useful to classify long and noisy time series [169]. Dictionary based methods fit into this category and have recently been used with great success. These techniques rely in a transformation of the time series into a symbolic feature vectors by means of a specific method, such as the SAX [103] or the *Symbolic Fourier Approximation* (SFA) [170]. The first approach proposed for time series classification with symbolic representations was the work of *Jessica Lin et al.* with the *Bag of Patterns* (*BoP*)[102]. Further proposed methods were conceptually inspired on the *BoP*, using the same reasoning. Techniques such as *Bag of SFA Symbols* (BOSS) and Word ExtrAction for time SEries cLassification (WEASEL), from the same authors, use a similar reasoning but employ the SFA instead [169, 171].

Using syntactic methods has already been successful for several time series data mining tasks, mostly related with query search and classification. Besides, these methods, being dictionary-based, can be used to show similarity between subsequences by looking into the distribution of word counts. However, current methods rely mostly in incomprehensible sets of characters, such as `aaa`, which are hard to associate with a specific subsequence of the time series, therefore providing limited interpretability. In this manuscript, we propose a method that translates the time series into sentences, such as that if a human was to describe a time series with text, it should be possible to separate these time series with the written words. We have seen natural language being used to include the human in the loop for more intuitive and meaningful query searches in time series [79, 80].

There is an existing method that is capable of providing visual interpretability of differences between time series, which is a structure-based method called *shapelets* [208]. Shapelets are representative subsequences of the time series, which characterize a specific class. The advantage of this method is the higher interpretability because relevant shapes from the class can be highlighted [208].

All the mentioned methods are a reference in time series classification tasks with innovative concepts that merge ideas from the text-mining domain. One of the advantages of structure-based methods that rely in a dictionary-based concepts is to use the words extracted as an interpretable model to differentiate time series. The histogram of words generated gives the user an understanding of which patterns better represent the time series and give an intuition over patterns that differ between classes of time series. This provides a feedback and explanation over why a class is different than the other. However, dictionaries can be confusing, and the words generated are not intuitively associated with the patterns these represent in the time series. One method that went beyond the previously mentioned methods in that aspect is the SAX-Vector Space Model (SAX-VSM)[174]. This method used a weighted word vector representation of the time series and showed which are the relevant words for the classification process and what patterns

these represent in the time series, demonstrating that the data can be interpretable by measuring the importance of the patterns found for each class of signals [174].

The proposed method is built upon the same ideas as the BoP method but uses the proposed SSTS method to promote the inclusion of the human reasoning in the classification process and provide more interpretable representations, as inspired by the work of SAX-VSM.

3.5 Applications in biosignals

We live in an "era of big data" [33]. As mentioned in Section 1.1, wearable sensors are currently available on a large scale, promoting the acquisition of massive amounts of data. Datasets of this size can no longer be handled by trivial means and call for engineers and data scientists with expertise in data mining, machine learning, and data analysis [200]. This increase in wearable usage has also been seen in industrial environments, which is motivated by the current trend of Industry 4.0 [203], promoting the use of sensors to monitor in real-time their machines for damage prevention, and their workers for occupational-related disorder prevention and productivity improvements [190, 166].

Research areas such as intelligent rehabilitation [108, 146, 24, 182, 30, 85], orthotics [214, 129, 213, 122], sports science [101, 109, 128, 87, 75, 127, 74, 211, 46, 144], activity modelling [104, 32, 31, 107], exoskeletons [199, 56], machine learning education [67], and fall detection [29, 141, 204], have also leveraged the power of biosignals from wearable sensors.

In research aspects of time series analysis, biosignals produced by various types of sensors require the data science community to develop tools to extract meaningful information for the acquisition, including reporting, pattern recognition, event detection and periodic signal segmentation and classification, among other data mining tasks [158, 21]. The availability of more reliable data and practical information is more beneficial, primarily as machine learning is increasingly applied. Numerous fields could benefit from the methods explored in this work, including physiological event detection for healthcare (e.g., noise, sleep problems and epilepsy), biomedical signal analysis (e.g., ECG, EEG and EMG), climate change detection, audio-based automatic speech segmentation and recognition, motion sequence segmentation, behaviour transition detection, human activity modelling, feature space study, and manufacturing industries. In that sense, we will present a general set of examples where the proposed methods can be applied with practical use. In a first instance, the proposed tools facilitate data scientists in exploring their data, but we hope that future work on these tools can even democratize their usage to other analysts.

4

Unveiling the *Grammar* of Time Series

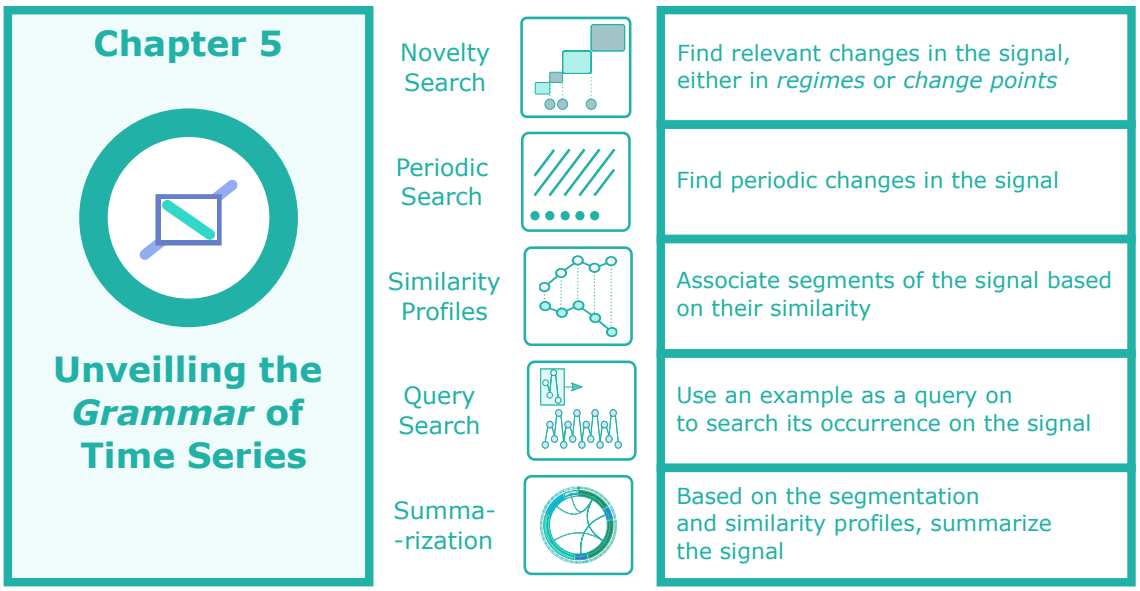
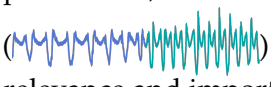


Figure 4.1: Information retrieval topics explained in this Chapter. Each of these topics is a result of analysing the SSM.

Researchers are interested in understanding the structure of the recorded signals, the meaning behind them, and the influences of the context. In Chapter 1 (see Section 1.2) we gave several examples of how signals in our everyday life can be structured and partitioned, such as a signal representing a walking regime shifting to a jogging regime (). The examples we mentioned in Chapter 1 and above manifest the relevance and importance of the following topics:

- **Novelty segmentation**: to identify significant changes in the signal's behaviour.
- **Periodic segmentation**: to detect the presence of repeating cyclic patterns.
- **Labelling**: to measure how similar the segments are between each other with the help of similarity profiles.

- **Query search**: to find repeating patterns in the time series based on a *subsequence* given as an example.

- **Summarization**: to summarize visually the time series by partitioning it into relevant and uniform/periodic segments and joining segments based on their similarity.

In this Chapter, we present the proposed solutions to the five problems mentioned above, which are inspired by a method used for audio signal analysis and thumbnail generation [133, 147, 20, 22]. Such a method has not yet been extended to other types of *time series* domains that could greatly benefit from it [5]. The method uses a feature-based SSM of (multidimensional) time series, from which visual and analytical information is rendered to perform the segmentation process and associate *subsequences* of the time series with each other. This association based on similarity can be used towards automatic labelling and ultimately to summarize the signal. We extended its application to *biosignals* with the usage of general features and introduced new methods to analyse the SSM for periodic segmentation, labelling, query search, and summarization.

4.1 The Problem

Defining what is relevant in a time series highly depends on the context and purpose of the analysis, but globally, for any type of time series, there is a general interest in understanding how the signal is structured and how are *subsequences* related, especially for tasks related to data labelling. The structure of a time series is built of *segments* delimited by *events*. The problem explored in this section is the search for such *events* and how to find the ones that are significant and useful to then perform a similarity search and summarization of the time series.

From the definition of *event* (see Chapter 2 Section 2.1), we highlight two primary considerations for the detection of events: (1) an event is a change in the behaviour of the time series, and (2) it has to be significant both *statistically* and *qualitatively*. The *qualitative* aspect indicates subjectivity from the analyst because of the domain or context of the problem. Considering this, we will start by explaining the dimensions of the problem: (1) search and (2) type of significance.

4.1.1 Search Ranks

Figure 4.2 illustrates the search ranks of the problem formed by three layers:

- **Dimensionality**: The search can be applied to one or multiple time series. In multidimensional space, some events can coincide in several time series, while others are specified on a particular dimension. For example, some gestures produce noticeable signals on only one dimension of the three-axis accelerometer.

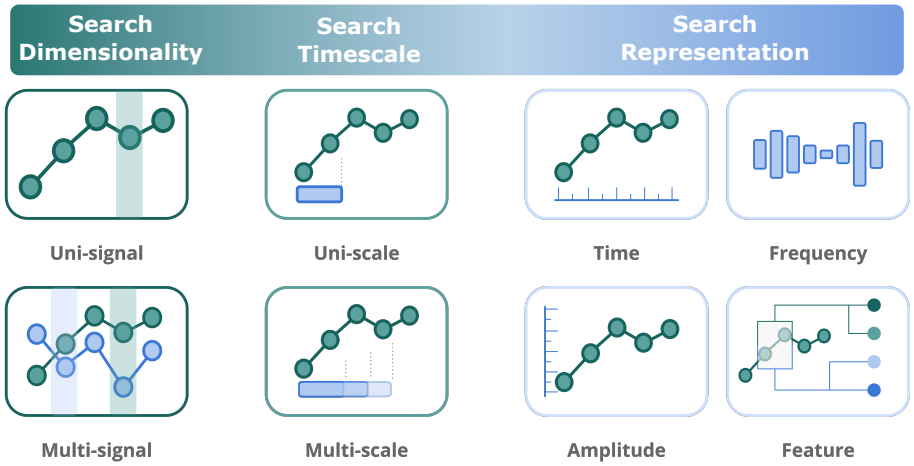
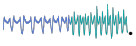


Figure 4.2: Event search in different ranks of dimensionality, timescales, and representation.

- **Timescale**: *Events'* occurrence can vary from different timescales. For example, when the signal being analysed is zoomed in from hour to minute scales, some events may disappear while new events may be detected.

- **Representation**: The searchable objects can be straightforwardly the temporal nature of time series or other representation, such as frequency or other extracted features.

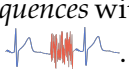
Besides the ranks mentioned above, the search procedure can be customized by context or target, which is highly related to the relevance given to an *event* or a *subsequence*. Types of events that are considered significant include:

- **Property change**: The change of a property, such as a change in mean or frequency, or a set of properties is greater than a threshold, e.g., low to high frequency -  .

- **Peak/Valley**: Peaks and valleys can typically be associated with significant physical changes, e.g., ECG peaks like  .

- **Periodicity**: The starting points of each period in a periodic signal are considered relevant, e.g., ABP periods like  .

- **Recurrent pattern**: Re-occurrences of similar *subsequences* with specific patterns should be of interest. Unlike *periodicity*, *recurrent* patterns do not have a temporal regularity, e.g., the arrhythmias found in an ECG signal like  .

- **Anomaly**: Highly dissimilar *subsequences* with particular patterns are of the reference value, e.g., noise in a clean signal  .

4.1.2 Proposal

In order to fill as many research gaps as possible, this study started by defining the search space for the aforementioned topics, considering that if the time series is transformed in the feature space, we can have a multidimensional similarity analysis based on the features' behaviour. The notion of *difference* in time series can be associated with *distance/similarity*, enabling finding segmentation points, recurrent patterns, anomalies, and periodic shapes, which is ideal to perform all the approaches listed above. This feature-based multidimensional representation is already successfully used for time series' classification purposes [16] and is potentially applicable to all the mentioned topics. For instance, any feature's change would be relevant for novelty segmentation, such that changes in the mean, standard deviation, frequency, or other properties are all options worth searching for. By characterizing the signal in the feature space, we can explore changes in all feature representations.

We propose an unsupervised methodology that searches for events (1) in uni and multidimensional space, (2) with a fixed timescale and potential multi-timescale application opportunities, and (3) on an SSM computed by a feature space representation of the time series. The events to be searched are any changes in the SSM related to a segmentation point and/or a periodic event.

The proposed method's reliability for event detection will be evidenced by considerable experiments in various type-agnostic databanks of multiple time series domains and comparisons to state-of-the-art methods. It should be highlighted that events in different datasets are extracted from the same information source, i.e., SSM, which meanwhile provides insights into unsupervised automatic labelling.

4.2 Building the Self-Similarity Matrix

In this section, we explain the steps of the proposed method. The extraction of relevant events from time series starts by computing the SSM. As explained in Section 2.6.2, this matrix has relevant structural information to retrieve *events*, namely *blocks*, *paths* and *similarity profiles*. Figure 4.3 summarizes the steps involved in calculating the SSM.

4.2.1 Feature Extraction

The structural information on the SSM reflects how informative the feature set can translate the signal's changes and disruptions. Behavioural changes may be related to a variate set of features. As a feature can be sensitive to a particular type of change, the set of features should be diverse to identify a multivariate set of events and be agnostic to various signal types. We turned to the available features from the *Time Series Feature Extraction Library* (TSFEL) [16] for Python, which has been proven effective and efficient in our previous work on Human Activity Recognition (HAR) and multimodal biosignal processing [105, 104]. We selected over 50% of all TSFEL features in the statistical, temporal, and frequency

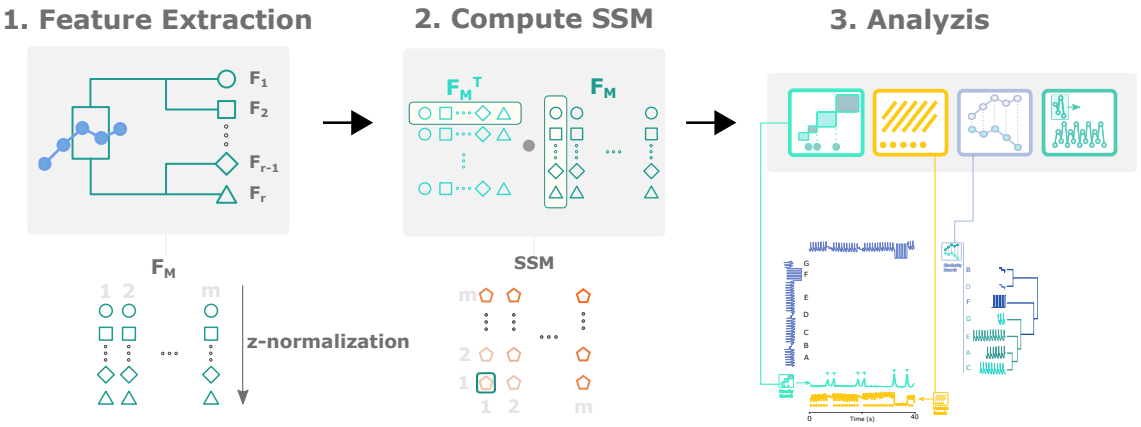


Figure 4.3: A step-by-step flowchart for calculating and analysing the SSM. The signal-based calculation requires input parameters of the window size w and the overlapping percentage o to fulfill the first-stage feature extraction. Features are extracted on each subsequence $(sT_1, sT_2..., sT_N)$, where N is the total number of windows. K features are extracted from window i (sT_i: $f_{i_1}, f_{i_2}, ..., f_{i_K}$). Different features are associated with different shapes (◯, □, ◇, and △) in the figures. The features can be extracted on an M-variable record and each feature is positioned as a row on the F_M for the SSM computation.

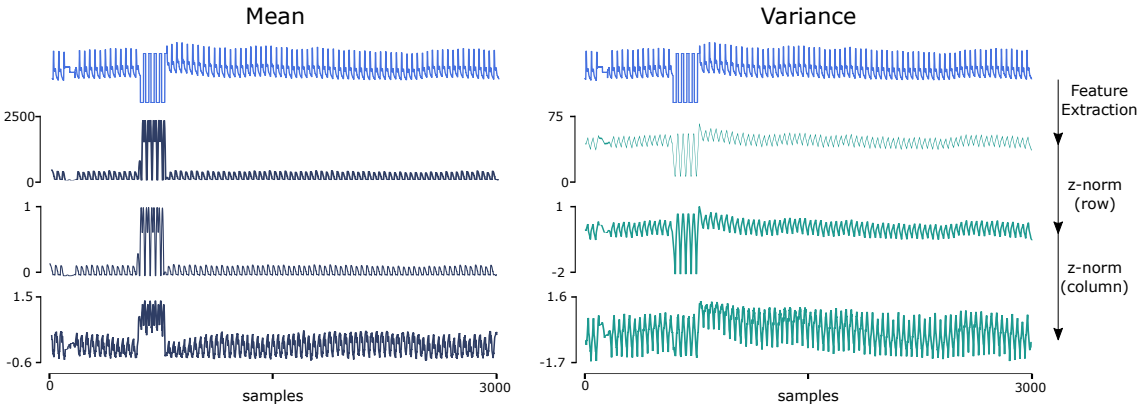


Figure 4.4: Example of feature vectors before and after normalization. Mean and Variance features are presented for the ABP signal from Dataset A.1.2.

domains with relatively lower computational costs, as listed in Table B.1 in Appendix B.1, regarding our proposed method's high calculation resource consumption.

The features are extracted with a moving window with size w, specified by the user, with an overlap percentage o. The selection of the two sizes significantly influences the results: w defines the timescale at which features are extracted so that the wider the window, the more *zoomed-out* the search will be. The second parameter defines the pixel resolution of the resulting feature series, increasing the amount of information with a larger overlap.

The extracted features are grouped into a feature matrix (F_M), where the rows represent a feature series and the columns correspond to all subsequences. In the multidimensional case, r features extracted from each of the k dimensions are ordered in the F_M as rows, forming $r \times k$ elements in each row, as illustrated in Figure 4.3.

Each feature series (rows of the F_M) is z-normalized for a more balanced contribution to characterizing the signal. A further normalization is applied to the feature vector (columns of the F_M), which optimizes the cosine similarity computation between feature vectors by simply adding the dot product to calculate the SSM. As an example of two simple features (mean and variance) and their normalized versions, we show Figure 4.4.

4.2.2 Feature-based SSM

After grouping all the features extracted, the next stage is to apply a similarity measure to the feature space and compute the SSM. This process consists in comparing each *subsequence* with all the other *subsequences*. Since each column of the F_M is each subsequence's feature characterization in the entire feature set, the SSM, i.e., the comparison between segments, is obtained by calculating the dot product between the z-normalized transposed F_M and itself:

$$\text{SSM} = F_M^T \cdot F_M. \tag{4.1}$$

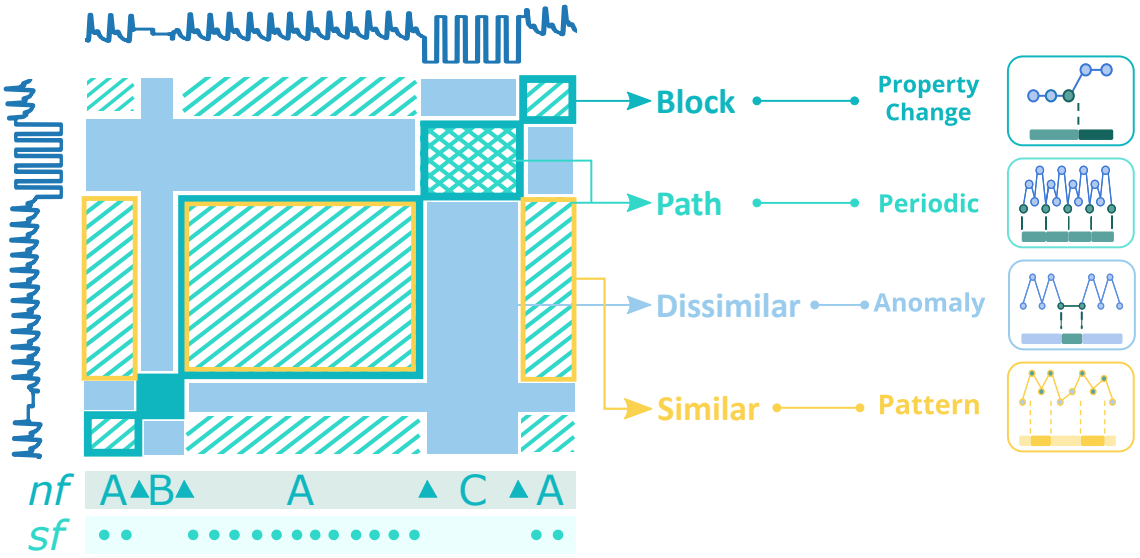
The dot product scores the similarity based on the subsequence's feature values. Cells of the SSM with higher similarity scores indicate that the corresponding *subsequences* have similar feature values [147, 20]. As a result, the SSM provides rich visual information, highlighting structures that describe the signal's morphological behaviour over time and structure, such as blocks and paths.

In Figure 4.5, the main structures are illustrated and highlighted in an example of an *SSM* [147] computed from an ABP signal, where the main structures are *blocks* and *paths*. Our proposed method utilizes the resulting main structures to extract the desired information. *Paths* show recurrence of patterns, which indicates the morphological matching between corresponding *subsequences*. Circles in the *sf* layer exhibit when the paths start. The *cross-pattern* in *block* C means that the *subsequences* are periodic and symmetric.

Differently, *blocks* are square-shaped structures of homogeneous areas in the SSM, translated as constant behaviour in the time series. The change between block structures along the main diagonal displays a relevant change in morphology and behaviour in the time series. In Figure 4.5, the SSM is segmented into several blocks on layer *nf*, for which the $\triangle$s mark the change points that separate blocks A, B and C. Besides *paths* and *blocks*, the SSM provides similarity measures between *subsequences*, which can be used to spotlight (dis)similar segments, such as anomalies, motifs or cycles. Several strategies were applied to the SSM to extract the mentioned information.

4.3 Information Retrieval

The SSM is a powerful visual tool *per se*, exposing relevant information that a raw observation could miss. Automatic discovery of information of interest will increase the SSM's



.

Figure 4.5: The informative structures of an ABP signal's SSM. The three main structures are highlighted in the simplified illustration: A - the homogeneous segments corresponding to periods in the ABP signal; B - the homogeneous segment representing missing data; C - the homogeneous segment cueing sensor detachment. The "blocks" in the figure accentuate homogeneous behaviour while the paths in the figure depicts periodicity in the segment. Segment C has a cross pattern, which symbolizes periodicity and symmetry. nf: *novelty function*; sf: *similarity function*; $\triangle$: change points separating blocks A, B and C.

practicability and versatility, for which three approaches for information retrieval on the SSM are put forward: (1) novelty search of *block* transitions, (2) periodic pattern search of *paths*, (3) similarity profiles of *subsequences*, and (4) how to use a query from the SSM to search for specific *subsequences*.

4.3.1 Novelty Search

The search for *novelty* is inspired by a method used in musical structure analysis [49], which is computed with the help of the *libfmp Python* package [135]. The process involves searching for transitions between *blocks* using a moving chequerboard square matrix, resulting in a one-dimensional function: the *novelty function*.

As shown in Figure 4.6, *block* transitions along the diagonal are represented by a chequerboard pattern. Such patterns can be detected by correlating a standard chequerboard matrix with the diagonal of the SSM, for which a sliding squared matrix, designated *kernel*, is used. The kernel incorporates a Gaussian function with a smoothing factor. The kernel K_N combines two different square matrices: K_H and K_C. K_H is responsible for identifying the homogeneity of the SSM on each side of the centre – the more homogeneous the pattern is, the higher the corresponding values will be. K_C measures the cross-similarity level. Therefore, when sliding the kernel K_N along the diagonal, a higher correlation value will be returned when it reaches a segment of the SSM with a similar chequerboard

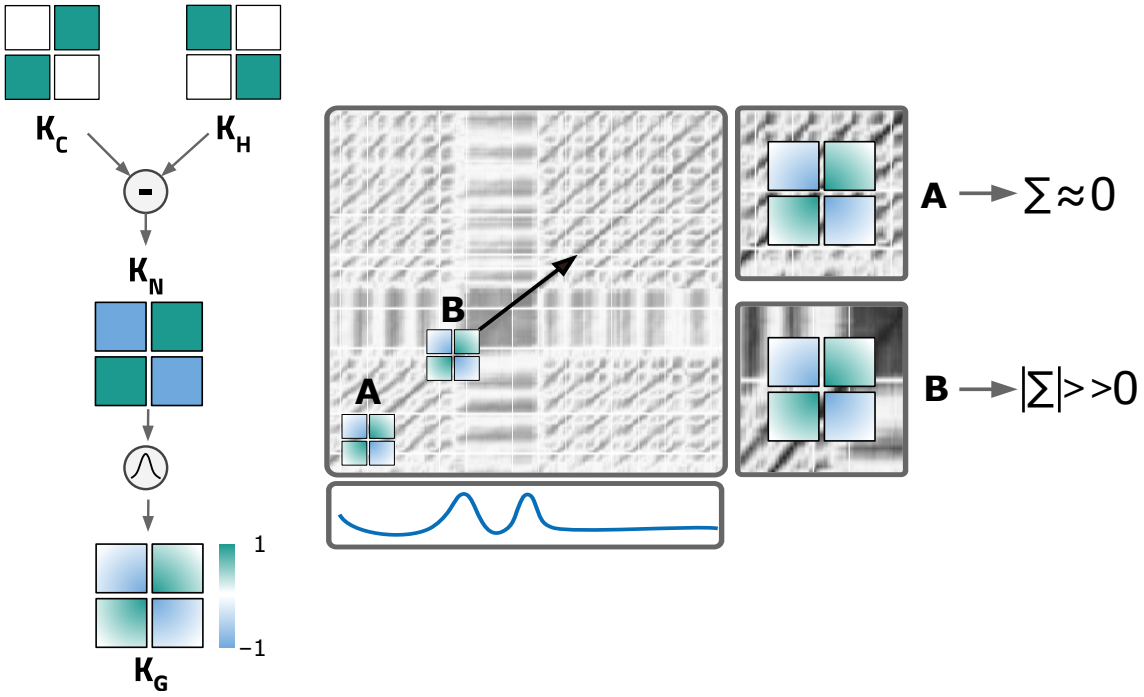


Figure 4.6: Left: description of the matrix (kernel) used to compute the *novelty function*, based on the works of *Mueller et al.* [133, 134]. The chequerboard pattern of the kernel K_N is achieved by combining the kernel K_H (homogeneity measure) and K_C (cross-similarity measure). Combined with a Gaussian function, the K_G is obtained; right: the process to compute the novelty function based on the works of [43, 133, 134]. Kernel K_G slides along the diagonal of the SSM to compute the *novelty function* presented as the bottom sub-plot. Positions A and B point to the effect of block transitions on the *novelty function*.

pattern. The result is the mentioned novelty function [43, 133, 134].

In position **A** of Figure 4.6(right), due to the high homogeneity, the kernel returns a value approaching 0 when summing the product between it and the section of the SSM it overlaps. In contrast, the kernel in position **B** reaches a segment with low cross-similarity and high diagonal similarity, which results in high correlation values with a chequerboard pattern. Therefore, high *novelty function* values are witnessed in these transition segments [43, 133, 134].

Each section of the kernel has the same size, $L \in \mathbb{N}$, and $D = 2 \times L + 1$ configures the total kernel size. The kernel has an odd size to adapt zero values in centred points, and a total size of $D \times D$. K_N is defined by [133, 134]:

$$K_N(i,j) = sign(a_i) \cdot sign(b_j), \tag{4.2}$$

where $a, b \in [-L : L]$ and $sign$ represents the sign function (1, 0 or -1). A radially symmetric Gaussian function is used to smooth the Kernel [133, 134]:

$$\phi(p,u) = \exp(-\frac{1}{2L\sigma^2}(p^2 + u^2)), \tag{4.3}$$

where σ is the standard deviation, equal for both x and y dimensions of the matrix, L the size of each kernel's section, and p and u the position in the x and y dimensions, respectively. The kernel K_G is computed by point-wise multiplication with the Gaussian function:

$$K_G = \phi \cdot K_N \tag{4.4}$$

The *novelty function* n_f is calculated by correlating the kernel with the diagonal of the matrix:

$$n_f(m) = \sum_{i,j=0}^{2L+1} K_G(a_i, b_j)\,\mathrm{SSM}(m + a_i, m + b_j) \tag{4.5}$$

being the sample of the novelty function $m \in [0 - N]$ and $a, b \in [-L : L]$. The change point events are represented by local maxima (peaks) in the *novelty function*, which can be detected by standard peak-finding strategies.

4.3.2 Periodic Search

As aforementioned, *paths* indicate the presence of similarity and reoccurring patterns can be visualized on the SSM. The *path*'s start point punches where the period of the pattern begins. In order to find the periodicity, we compute the similarity function s_f by summing the values of the symmetric SSM column-wise or, equally, row-wise. Each element of the sf is calculated by

$$sf(x) = \sum_{i=0}^{m} \mathrm{SSM}_{ix}, \tag{4.6}$$

where i is the column position for the sum, sf_j is the sample of the function at position j and m is the feature-series size. As segments with similar morphology will be closely described by the extracted features, the columns will have an approximate representation, resulting in similar values on the *sf*. The similarity function will enhance such behaviour when facing periodic series. The identification of events related to the periodicity of a time series is then feasible by searching for local minima (valleys) on the similarity function.

Although not validated in this work, an additional application of the similarity function should be outlooked. Considering that each sample of the sf is an average similarity of a subsequence to all other subsequences, it is possible to find *anomalies*. Regarding an *anomaly* as a subsequence highly unique and different from all the rest of the time series, its average similarity to all the other subsequences should have a low value.

4.3.3 Similarity Profiles

The principal elements, *blocks* and *paths*, are the information basis for segmenting the time series. Besides, SSM also provides pairwise similarity values between all *subsequences* of

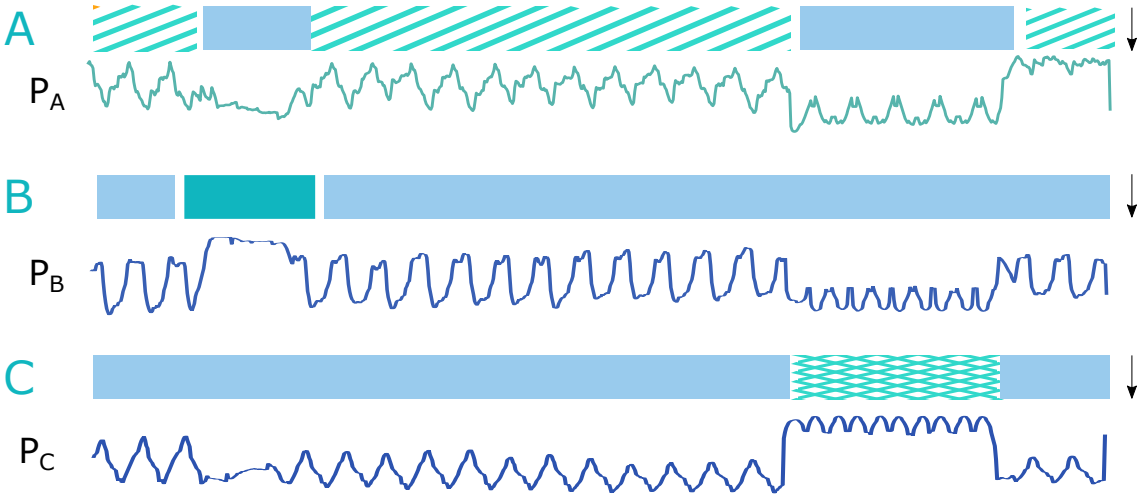


Figure 4.7: Profiles computed for each segment of the example signal used in Figure 4.5.

the time series, an important measure that can be used for clustering and *motifs*/*discords* discovery. The similarity profiles exploit the similarity values of the SSM to facilitate *subsequences* comparison. A similarity profile charts the similarity values of a *subsequence* (one column/row of the SSM) to all the other *subsequences*. The higher the values, the more similar the *subsequences*. In addition to the *subsequences* comparison, the similarity profile can also compare between entire segments of the signal. Consider, for instance, all three A-segments highlighted in Figure 4.5, whose profiles are highly similar despite the different sizes.

Although the segment comparison could be directly based on the region of the SSM delimited by two *subsequences*, we propose a more effective measure of two segments' similarity/difference according to their similarities/differences to all the other *subsequences*. A *similarity profile* $P_s(c)$ of a segment is computed as the column(row)-wise average similarity values of the region delimited by the segment being profiled (size l), and all the other *subsequences* of the time series (size m):

$$P_s(c) = \frac{\Sigma_{i=0}^{l} SSM(i, c)}{l}. \tag{4.7}$$

The *similarity profile* is computed column(row)-wise. Each column(row) $c(r)$ is the average similarity value between the reference segment and the segment corresponding to c. The reasoning is that similar segments should have closer *similarity profiles*. Since the profiles have the same size, they can be compared by certain distance measures, such as the ED, to form clusters, an automatic clustering solution based on the segments generated by the *novelty* and *similarity* functions.

A general example of applying this process to the segmented time series based on the *novelty function* is shown in Figure 4.7. Each segment category (A, B and C) extracted from the SSM of Figure 4.5 is computed into a profile (P_A, P_B and P_C) by averaging column-wise. These *similarity profiles* show how similar the segment is with all the other *subsequences*

of the time series. All segments A will have a similar P_A while being very different from profiles P_B and P_C.

4.3.4 Query Search

The mentioned similarity profiles are also useful to search for specific repetitions of a query from the time series. The process follows the traditional methods of template-based search methods explained in Chapter 2.6.3, but instead of doing the process directly on the time series, it is performed on the SSM. The search procedure works by sliding the smaller column window (the example selected on the time series) along the SSM, one sample at a time. The distance, D, between the example and the segment it slides over is calculated through the sum of absolute differences:

$$D(i) = \sum_{i=0}^{i=m} \sqrt{(\mathrm{SSM}(i) - \mathrm{SSM}_t)^2}, \tag{4.8}$$

where $\mathrm{SSM}(i)$ is the segment of the SSM over which the example, SSM_t, slides at moment i, up to the size of the SSM. The resulting distance function has minimums at the position where the example is matched.

4.4 Illustrative Evaluation in Various Application Scenarios

Experiments from multiple domains were carried on to validate the practicability and universality of the process to represent the time series into a feature-based SSM and the methods of retrieving information from the SSM.

4.4.1 Acceleration Signals in Human Activity Domain

Figure 4.8(top) exemplifies the SSM's usage on a record of an HAR dataset (*HAR1*, see Section A.1.3), where the data of all three accelerometer axes is applied. The SSM was computed using a 250-sample window size, and a 95% overlap. Along the diagonal, the novelty function generates block-wise references for estimating activity transition using a 45-sample kernel.

We can identify in Figure 4.8(top) that the detected segmentation points match the activity transitions. Although all transitions are visible on the novelty function, the transitions between similar activity patterns in the walking category (straightforward, upstairs, and downstairs) are more challenging to differentiate, as block A suggests, which is plausible since the properties of these segments are morphologically similar.

The proposed unsupervised method automatically and sensitively detects any significant change in properties. As can be found in the yellow-marked part in Figure 4.8(top), the period in which the subject was performing the *Upstairs* activity is affected by other changes in the time series. These are significant and also correspond to *block* transitions, which are also evident in the novelty function.

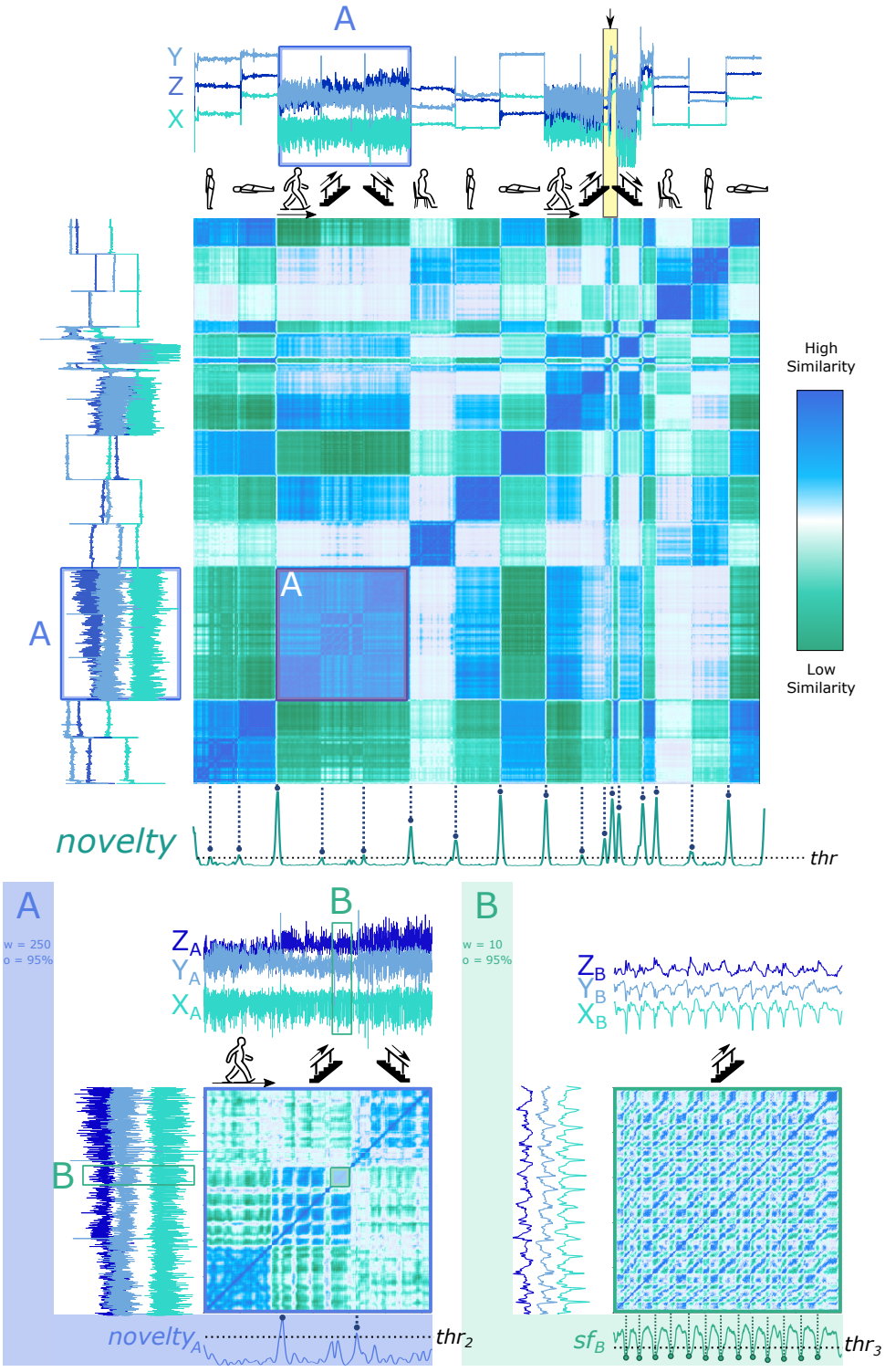


Figure 4.8: An SSM-based novelty search strategy to detect segmentation events on a signal piece from Dataset 3 (see Section A.1.3). Top: $window_{size}$=250 samples, $kernel_{size}$=45 samples, and overlap=95% on the activity sequence Standing → Laying → Walking → Upstairs → Downstairs → Sitting → Standing → Laying → Walking → Upstairs → Downstairs → Sitting → Standing → Laying; bottom left: signal change point detection on segment *A* with a size of 5000 samples, an overlap of 75%, and a kernel size of 25 samples; bottom right: further zooming in with a window size of 10 samples and an overlap of 95%, to reveal more periodic details of segment *B*.

When *zooming* the SSM into segment A in Figure 4.8(top), the three activities in the walking category can be effortlessly segmented based on the change points revealed by chequerboard patterns, as the two most prominent peaks in he corresponding novelty function pinpoint (see Figure 4.8(bottom left)). In addition, it is noticeable that the matrix segments related to *Upstairs/Downstairs* are also segmentable into smaller *blocks*. As the information is not available in the dataset description, we believe they are a flight of stairs.

Questions may arise at this point. Why is the signal periodicity of the three walking activities not evident in Figure 4.8(bottom left)? The reason is that the window size used to compute the SSM is relatively large. If features are extracted with a smaller window size closer to the walking period, the *paths* delineating the pattern recurrence will be visible. Figure 4.8(bottom right) shows the SSM built from segment B of the original time series, with a window size of 10 samples and an overlap of 95%. The *paths* in the matrix enable the periodicity detection with the similarity function sf_B.

4.4.2 Arterial Blood Pressure (ABP) Signals in Posture Recognition domain

Many biomedical signals, such as ECG, ABP, and Respiratory Inductance Pletismography (RESP), contain retrievable structural information like periodicities. Meanwhile, unexpected changes may occur during the acquisition due to physiological responses, medical disorders, or sensing problems like noise, interferences, artefacts, and electrode detachment. We visualize two examples of physiological changes in different types of periodic signals.

The ABP signal can vary due to postural changes, as an available experiment at *Physionet* confirms [68, 61]. Figure 4.9(top) shows the process of segmenting the ABP signal based on postural changes, where the trapezoidal and the square wave tag the ground truth of slow and fast postural transitions. The proposed strategy well perceives the change points. Observably, the shape of the raw ABP signal in each regime is undistinguishable through the naked eye. Therefore, it is constructive to rely solely on the signal itself to implement postural change detection. It is important to point out that the periodicity of the signal is not visible on the matrix because the features were extracted with a window size of 5,000 samples, which is much larger than the period size. A smaller window size of 250 samples in the current scenario allows periodic segmentation, as Figure 4.9(bottom) illustrates, where the SSM is computed on the first 10,000 samples of Dataset 2 (*TILT*, see Section A.1.2). The resulting *similarity* function gives prominence to the periodic nature of the ABP signal.

The SSM of Figure 4.9(top) also shows which segments are similar to each other. The blue-coloured parts in the matrix indicate high similarity, presenting that segments from the same posture are more similar than between different postures. For further illustrative purposes, we computed the similarity profiles of each segment as if segmented by the *novelty* function, which evidences that the corresponding sections could be well clustered based on the similarity profiles ($P_A = P_C$ and $P_B = P_D$). In the same way, the similarity

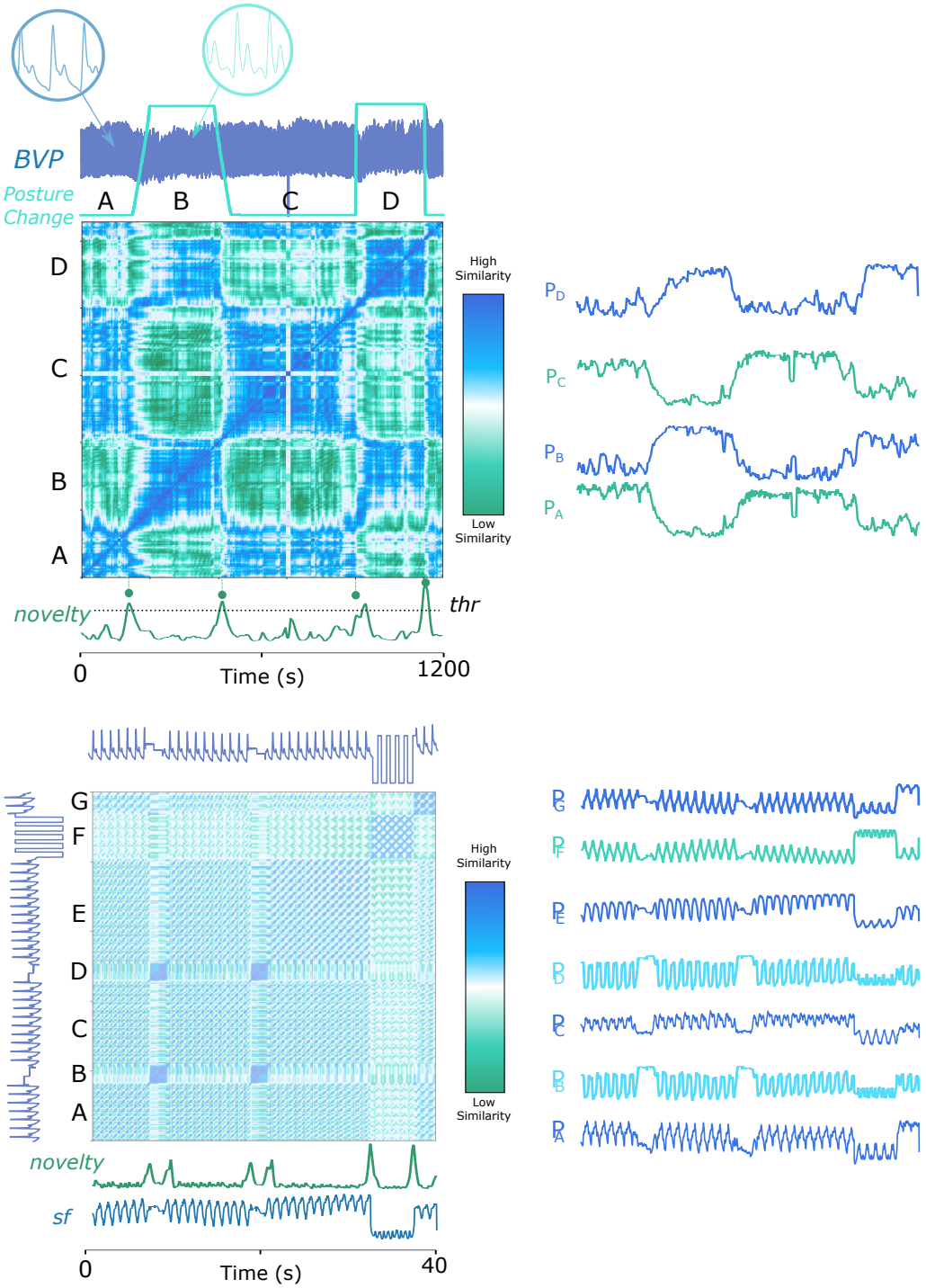


Figure 4.9: Novelty search on an ABP signal from Dataset 5 (*TILT*, see Section A.1.2). Top: a window size of 5,000 samples, an overlap of 95%, and a kernel size of 200 samples. The trapezoidal and the square wave mark the ground truth of slow and fast postural transitions. Bottom: the first 10,000 samples, with a window size of 250 samples, an overlap of 95%, and a kernel size of 200 samples. The right parts of the top and bottom subfigures plot the corresponding similarity profiles for each subsequence segmented by the novelty function.

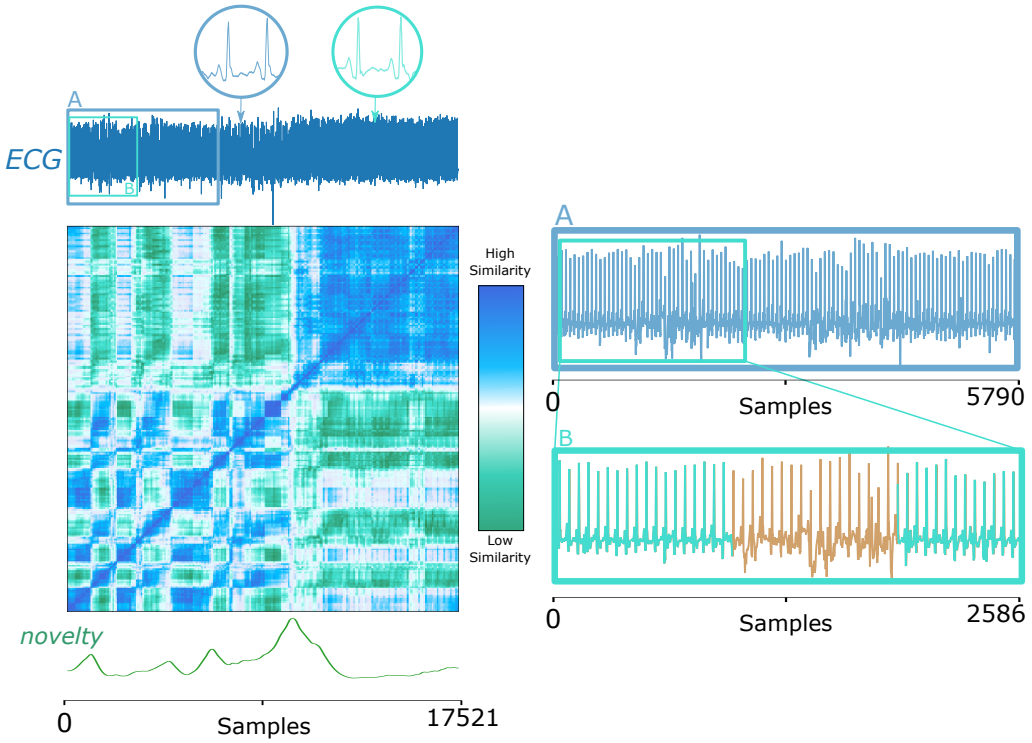


Figure 4.10: An ECG signal with a *pulsus paradoxus* condition starting at the $10,000^{th}$ sample from Dataset 6 (see Section A.1.14). Left: the SSM diagnoses two modes in the signal, whose patterns are zoomed in the circle thumbnails, respectively; right: zooming parts of the original signal can verify SSM's ability of automatic ECG pattern change detection and contribution to segmentation.

profiles in Figure 4.9(bottom) examine the similarity between segmented *subsequences*. Profiles with a similar shape can be grouped ($P_A = P_C = P_E = P_G$ and $P_B = P_D$), which can be applied to automatic clustering, as will be exhibited further in the next Section.

4.4.3 Electrocardiography (ECG) Signals in Biomedical Domain

Another widely used biomedical signal, ECG, also testifies to the feasibility of our proposed method. The ECG signal in Figure 4.10(left) displays the presence of the condition called *pulsus paradoxus*, an exaggerated fall (>10 mmHg) in the subject's blood pressure during inspiration [130], which can also occur when the patient changes sleeping posture after heart surgery[57], as the following example relates. Similar to the ABP signal elucidated in Section A.1.2, the human eye hardly perceives the change points in ECG signals. Once again, our proposed strategy shows strength.

In addition to the novelty detection, segment $\boldsymbol{A}$ previous to the pulsus paradoxus occurrence can be partially detailed again to reveal minor changes due to additional noise, verifying SSM's sensitivity to structural changes, as Figure 4.10 (right) imparts.

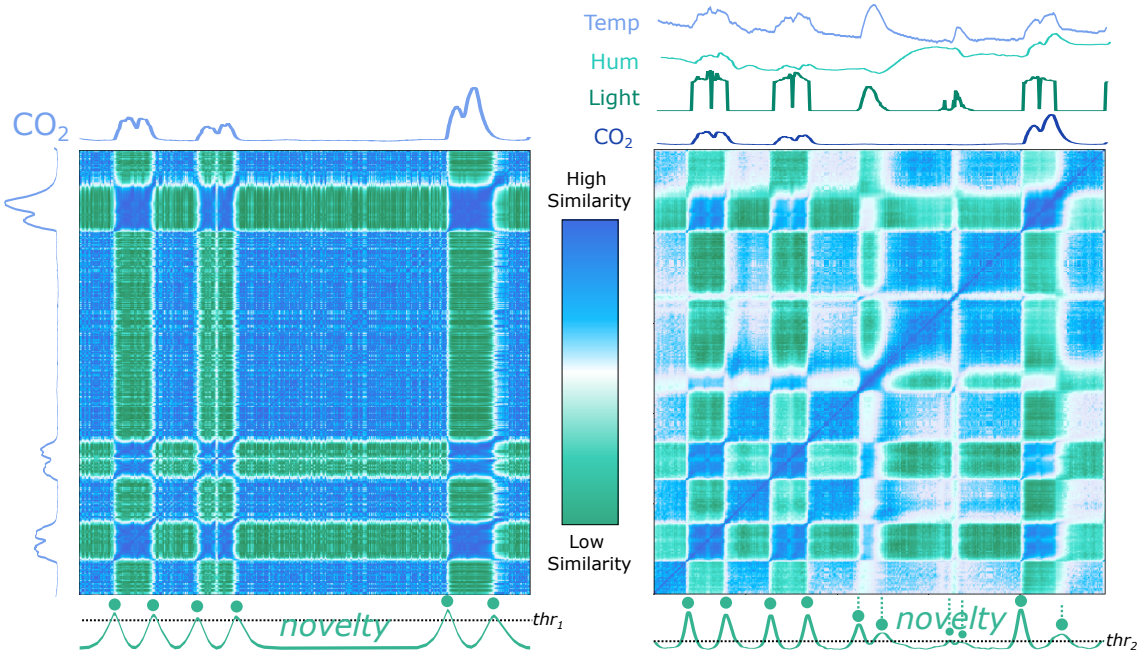


Figure 4.11: The proposed method applied on the *Occupancy* record of Dataset 9 (*ATCPD*, see Section A.1.9). Left: calculations on the separate CO_2 time series only; right: calculations performed by extracting features on the complete four time series.

4.4.4 Single Channel versus Multidimensionality Application in Multi-Sensor Scenarios

The proposed method accepts both single- and multidimensional records. The difference regards the number of features extracted. As compared in Figure 4.11, the same set of features is extracted from each time series to build the F_M. Using a single or several time series of a multidimensional record is an option, depending on the purpose. In some cases, the use of non-complete dimensions may miss relevant events, as Figure 4.11 instances the record "Occupancy" from Dataset 9 (*ATCPD*, see Section A.1.9).

The record is a multidimensional time series that measures room occupancy based on temperature, humidity, light, and CO_2. By comparing the signals and formations in the left and the right parts of Figure 4.11, it can be understood that some events can be detected using the CO_2 series exclusively, but some are missed.

4.5 Time Series Summarization

The aforementioned examples show how the proposed method can be used as a strong and reliable visual tool. It is possible to see how a time series is structured, how similar are segments, and if these are periodic or not. The information available is relevant to support the labeling process of the analyst, but it also can be used to summarize a time series and give a meaningful report about it.

Following the presented work, we studied how to use it for a meaningful summary

of time series. This process is inspired by methodologies that exist in other scenarios for data summarization techniques with statistical analysis, such as the available methods from the *pandas python library*: *pandas.profile()* and *pandas.describe()*. These methods can provide a summarization of a dataset (typically of categorical data) that is given as input. A similar method is not known for time series. In order to develop such a method, we should first understand what is meaningful and relevant to represent as a summary.

4.5.1 Elements with Relevance

In this section, we have been discussing which elements are relevant in time series, mostly associated with *events*. From *events* we can segment homogeneous *subsequences*, recurrent or periodic patterns and anomalies. The relationship between these segments is possible by analysing their similarity. In addition, a characterization of the segments is possible with statistical analysis. In that sense, the relevant elements to summarize in a time series are (1) *homogeneous segments*, (2) *periodic patterns*, (3) *recurrent patterns*, (4) *anomalies*, (5) association based on similarity and (6) statistical characterization.

4.5.2 Compact Design

Strategies that are typically used to present information compactly are found in several domains. In text analysis, for instance, the relationship between repeating sequences is illustrated with arc diagrams [96, 196]. These show where repeating sequences occur in a very concise way. This has a range of applications that include, for example, text and *DNA* sequence analysis. For instance, graphical genome maps are found to concatenate a significant amount of information in a very compact way. Genome features and sequence characteristics are assessed with this visual strategy. An example can be found in Figure 3.2b (see Section 3.1.4). This type of visualization inspired the summarization approach proposed for time series.

The proposed visualization strategy has several elements that can be used to transform the SSM into a compact form of filtered information. The elements are illustrated in Figure 4.12 and listed as follows:

- Novelty (N) and Subsegment Layers (P) - The novelty layer N has the segments resulting from the novelty segmentation step, color-coded based on how similar each *subsequence* is with the first segment. The subsegment layer P partitions each novelty segment into single periods, when the *subsequence* is periodic;

- *Chords* (C) - Each novelty segment is connected to the most similar subsegment with a colored connector;

- Signal Layer - The original signal is located on top of the partitioned segments.

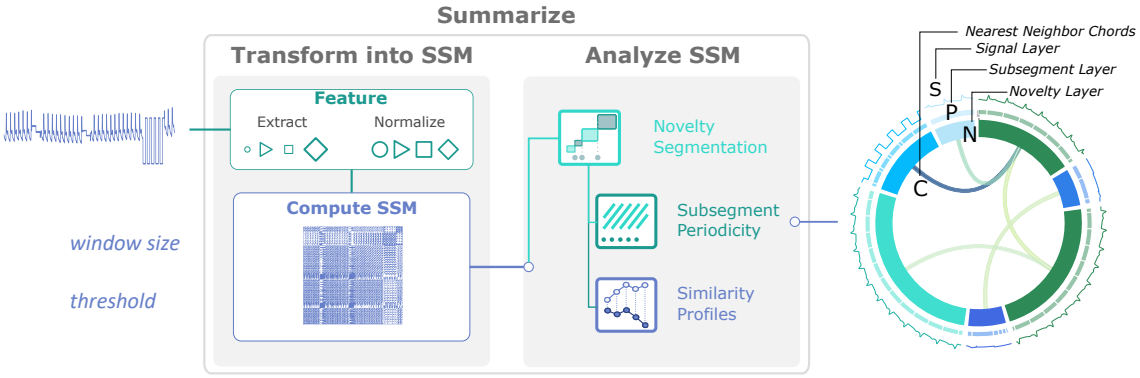


Figure 4.12: Steps to use the aforementioned methods on the SSM to summarize the time series into a circular plot with C - chords that connect nearest neighbors, S - the signal layer, P - the subseqment layer and N - the color coded novelty layer. Each of these elements of the summarized plots are built based on novelty segmentation, subsegment periodicity check and similarity profiles.

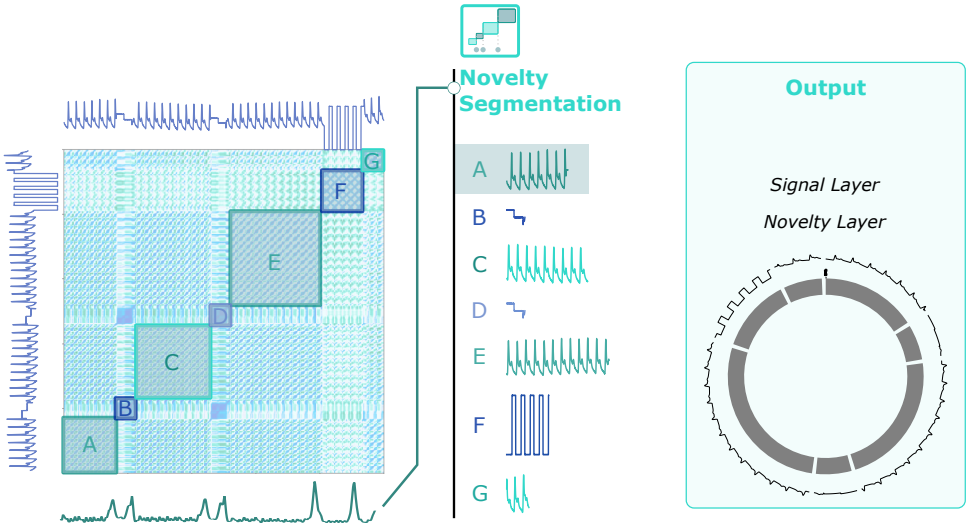


Figure 4.13: How to reach a standard format. This step applies the novelty search method to find the segmentation points. The signal is broken into A to G segments.

4.5.3 A step-by-step example

The steps are indicated in Figure 4.12 for the ABP signal example. After computing the SSM, it is firstly analysed to segment the signal based on the *nova* function. These segments are then compared based on the *similarity profiles*. Additional layers can be created by performing an iterative and multi-scale segmentation. With this process, the time series is segmented (*novelty layer*), subsegmented (*subsegment layer*), each segment is connected to the nearest neighbor segment (*nearest neighbor chords*) and the colors for each segment is given based on their similarity to the first segment. Figures 4.13, 4.14 and 4.15 show the step-by-step process to summarize the time series by analysing the SSM.

The *nova* function segments the time series into seven segments. The reader can notice that segments A, C, E, and G are similar and separated by segments B, C and F, represented

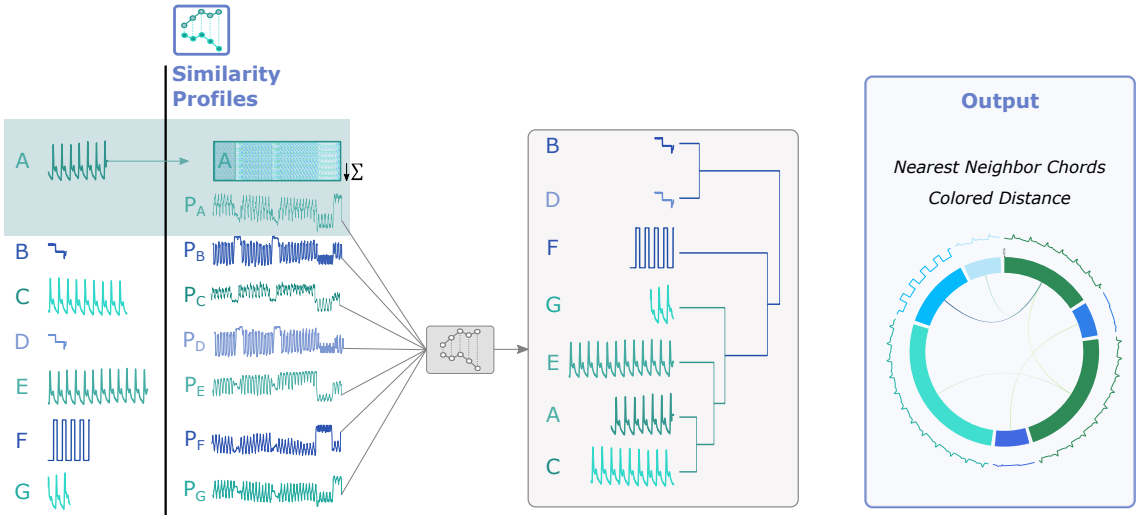


Figure 4.14: From each segment, a similarity profile can be computed. The pairwise euclidean distance is applied to these profiles, from which a dendogram is created. From this association, layer C can be drawn and the colors of the novelty layer can be added. Segment A is highlighted for the next step.

by a failure in the connection of the sensor. From this first segmentation, the *novelty layer* is created, indicating how structured is the signal, as presented in Figure 4.13.

Further, the segments are compared with the *similarity profiles* of each segment. Figure 4.14 illustrates as example the rows of the SSM delimited by segment A and the column-wise average into P_A. The same process is applied to each segment. From this Figure, the reader may notice that the profiles P_A, P_C, P_E, and P_G are more similar, while profiles P_B and P_D are more similar as well. The pairwise distance is computed between profiles, which can then be used to extract the nearest neighbor of each segment, as well as transform the distance values between segments into color.

The pairwise distance between profiles is also used to illustrate how the segments are ordered and clustered by the dendrogram of Figure 4.14. The dendrogram shows that there are three main clusters, being cluster one represented by segments A, C, E, and G; the second cluster has segment F and the third cluster groups segments B and D. It is important to note that typical distance measures, such as the ED or DTW, would not be able to directly sort these segments correctly. Using *similarity profiles* is more robust and invariant to the size and time distortions. The illustrated dendogram also has distance measures between the novelty segments. We use the distance from each segment to the first segment (A) and select the corresponding color-code to be represented on the summarized layout.

Finally, the process to summarize the time series can be iterative by adding *subsegment layers*. These layers can be added by performing the *novelty search* on the previously segmented time series with a smaller time scale or segmenting the time series based on periodicity. In this case, the signal is periodic, therefore Figure 4.15 illustrates the *periodic search*, where periods are segmented by the minima of the s_f.

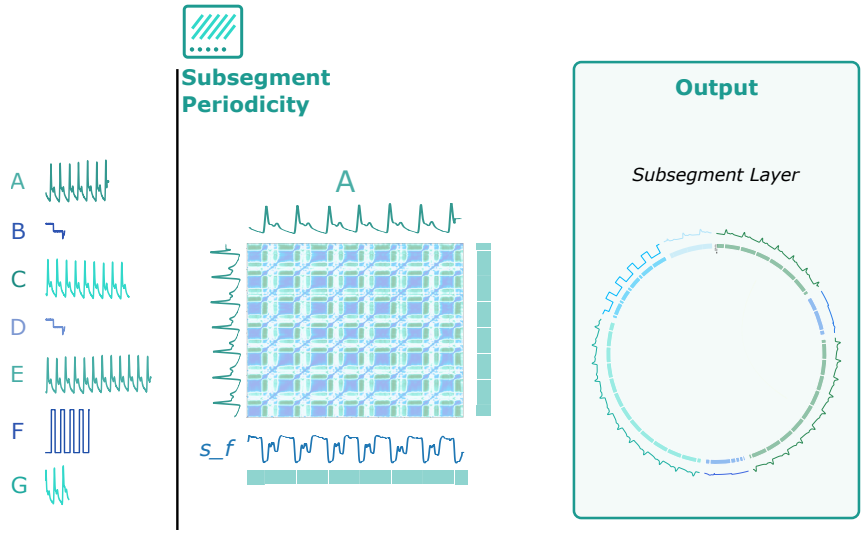


Figure 4.15: After finding the first layer of the segmentation points, the periodicity of each segment can be checked and added as a sublayer of segments, adding layer P. In this case, segment A is the example.

This summarization process is mostly graphical and follows the results of each process for information retrieval from the SSM. Therefore, this method will not present overall results or be validated in Chapter 6.

5

Language for Time Series Data Mining

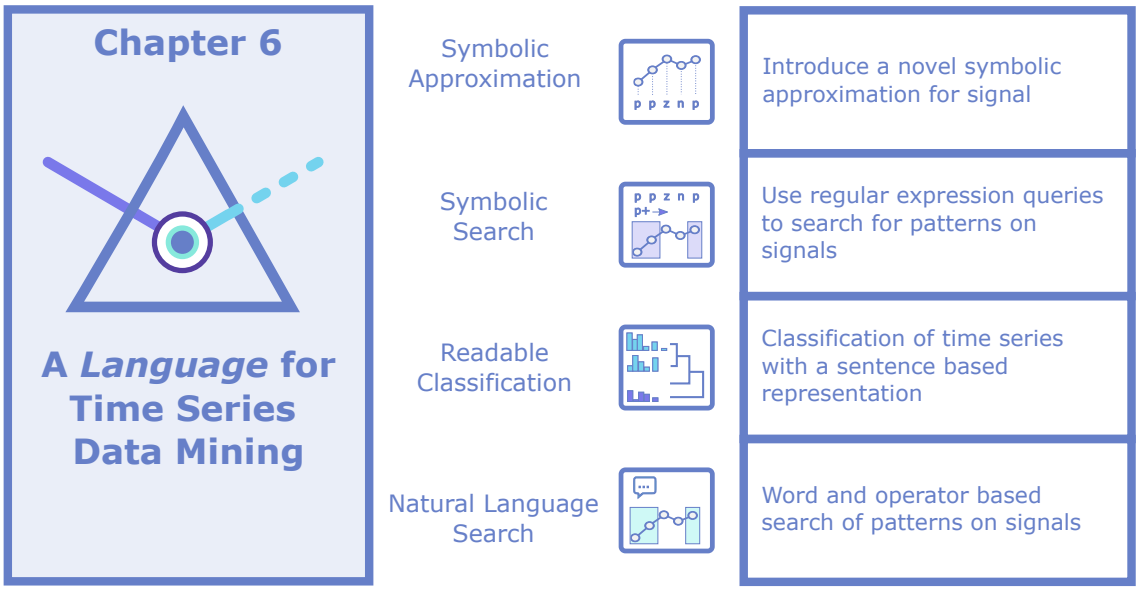


Figure 5.1: Four main topics will be covered in this Chapter, namely the novel symbolic approximation and how to use it for pattern search and classification. For these, we will introduce SSTS and HeaRTS. Finally, the search based on words and operators, close to natural language, will also be presented, introducing QuoTS.

Human language is foremost a means of communication, in which the information is represented by sentences, composed of words that can then be broken into sequences of symbols. Time series are, in their turn, carriers of information about a physical measure, represented as a sequence of ordered real domain numerical data observed during a given temporal interval. Analysts can have a visual perception of the morphological behaviour of time series and can describe it in such a way that another person understands which segment of the time series he/she is referring to. As we mentioned in Chapter 1, a physician may say "*I am searching for the T-wave, that represents the* `large peak`" ( ) or "*I am searching for the QRS complex, that looks like a* `sharp peak followed by a sharp valley` ( )". From this textual description of patterns, one can associate words with specific properties, such as `peak` can be related with *slope* and `high` related with *amplitude*.

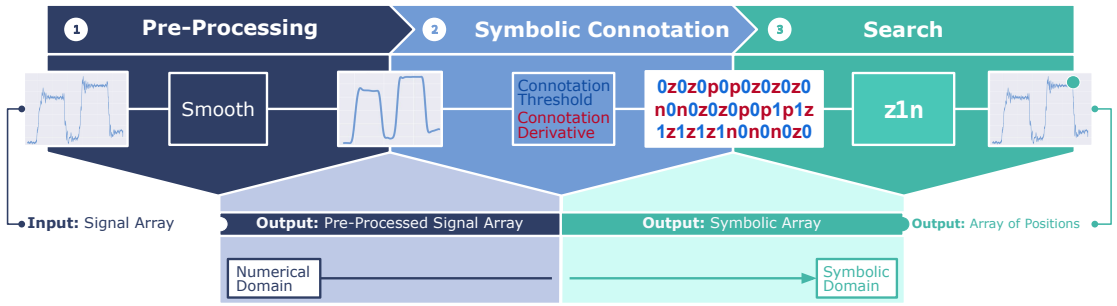


Figure 5.2: SSTS modular architecture: the proposed system is divided into three main modules - pre-processing, symbolic connotation, and search. This Figure also provides an example of the sequence of changes that occur in each step.

The text mining domain already has available well known approaches to search for specific words or patterns with text queries in traditional text editing softwares. regex are an example of such querying mechanisms. Other existing mechanisms rely in keyword-based queries, such as the *Google* search toolbar, which sorts web-pages according to how these match the set of keywords written. These are examples of the flexibility we find in tools available for query-based text search mechanisms. We imagined such tools could also be beneficial for time series query-based pattern search tasks.

In this chapter, we provide evidence that it is possible to use such reasoning with time series and propose solutions that promote higher flexibility, cognitive speed, and expressiveness in searching for patterns in time series as we would typically do with text. As presented in Figure 5.1, two main strategies are proposed: (1) Synthatic Search on Time Series (SSTS) and (2) *Quo -where-* On Time Series (QuoTS). The first explains how to perform query-based search on a symbolic representation of the signal. A novel symbolic representation is introduced and the usage of regex is proposed to search for patterns directly in this symbolic representation. We also show how this representation can be used in time series classification tasks with Human Readable Time Series (HeaRTS). The second method that is proposed uses a feature-based representation of the signal, being each feature assigned to one word, which can be used with additional *operators* to search for the desired patterns with natural language.

The chapter will start by detailing how SSTS works, more specifically the symbolic conversion and the search process. In addition, HeaRTS will be explained, namely how to use SSTS and its symbolic representation to find differences between time series and ultimately be used towards interpretable classification tasks. Finally, this chapter ends with QuoTS and how the analyst can use feature-words (keywords) with operators to search for patterns on time series more expressively.

5.1 Syntactic Search on Time Series

In Chapter 2 was presented the well-known SAX, which is a symbolic representation of the time series based on its amplitude distributions. With SAX, query with text

was not considered, but SSTS takes inspiration from this symbolic transformation by including additional attributes besides amplitude, such as the derivative, speed of the derivative, etc... (as will be listed further). Each property characterizes each sample of the signal. The combination of these attributes into a sequence of primitives is a symbolic characterization of the signal that enables the use of text-based querying mechanisms, such as regex, to search the desired patterns. Overall, the proposed tool benefits from the visual interpretation made by the analyst and provides SSTS to express it as a regex.

The proposed tool, SSTS, is structured in three stages:

- **pre-processing** - prepare the signal to highlight the desired patterns and ease the search process;
- **connotation** - a symbolic representation of time series into a set of primitives that give information about the shape and attributes of the time series over time and is governed by a set of grammatical rules;
- **search** - a parser (regex) to search for patterns in the symbolic representation. Both *pre-processing* and *connotation* steps rely on *tokens* to call specific methods.

The next sections explain each step of the SSTS process and will be accompanied by an example to elucidate how it works. The represented signal is the z axis of an ACC placed on the wrist of a subject while performing a rotating task at different angles. The purpose of this example is to find the time intervals where the plateaus above a certain threshold begin to decrease. Each step of this problem's resolution is shown in Figure 5.2 and will be thoroughly explained throughout the next sections. We take the liberty to start introducing some of the processing tokens (Sm is the smooth function, A is the Connotation Threshold and $D1$ is the Connotation Derivative) to give a base complete example of the usage of the SSTS.

5.1.1 Pre-Processing - Preparing the Data

The *pre-processing* stage is responsible for preparing the signal, either by removing noise that arises from several sources or adjusting the signal so that the information is unveiled. Traditional pre-processing methods, such as a pipeline of linear filters, moving window average filters and statistical de-noising or re-sampling techniques, are typical procedures to prepare the signal for further tasks [117]. The current approach uses a symbolic representation of these techniques, in which each is represented by its corresponding *token*, which can be a symbol or a function name. In order to manage the *pre-processing* tasks, a string containing a set of tokens and their corresponding arguments is written by following the polish notation [63], in which the token precedes the corresponding argument(s), and each element is separated by a white-space character. The available pre-processing methods to be used with SSTS are presented in Table 5.1.

Table 5.1: List of common SSTS pre-processing operators. As input parameters, s is the signal, *fc* is the cut-off frequency and *win_size* is the size of the window used (number of samples). The linear filters (HP, BP, and LP) have a default order of 2

Name	Input Parameters	Description
HP	(*fc*)	Linear high-pass filter with cut-off frequency *fc*
BP	(*fc1*, *fc2*)	Linear band-pass filter with cut-off frequencies *fc1* and *fc2*
LP	(*fc*)	Linear low-pass filter with cut-off frequency *fc*
abs	none	Modulus of signal - $\|T\|$
Mag	none	Magnitude - $\|T\| = \sqrt{T_0^2 + T_1^2 + T_2^2}$
Sm	(*win_size*)	smoothing of T by a moving average window filtering technique of size *win_size*
Z_{norm}	(*s*)	z-normalize the time series $\overline{T} = \frac{T - \bar{\mu_T}}{\sigma_T}$, being: $\overline{T}$ the normalized signal, $\bar{\mu_T}$ the signal's mean and σ_T the standard deviation of the signal
|	N.A.	Separates the pre-processing methods applied to multiple signals or multiple processing of the same signal

In the example of Figure 5.2, the *pre-processing* step uses the following string: Sm `500`, which indicates that a smooth filter using a sliding window with a size of *500* samples is applied to the signal. The resulting signal is shown under its original form. The selected area in the Figure 5.2 represents the segment of the signal that will be represented in subsection 5.1.2.

5.1.2 Connotation - The Symbolic Time Series

In semiotics, it is stated that the connotative aspect of an image or a sign corresponds to the extraction of meaning (sometimes called the *second meaning*), by the personal interpretation of its traits and characteristics [202]. The *connotation* step is the same connotative aspect but in this case, instead of a regular image, it is a time series. We follow the reasoning that the interpreter (the analyst) has made his analysis of the time series and retrieved the necessary information to search for the desired patterns.

The symbolic *connotation* step generates a sequence of *characters* by extracting properties of the signal that are based on a conversion rule that the user defines. Each of these conversion rules is designated a *connotation* method. These are responsible for converting

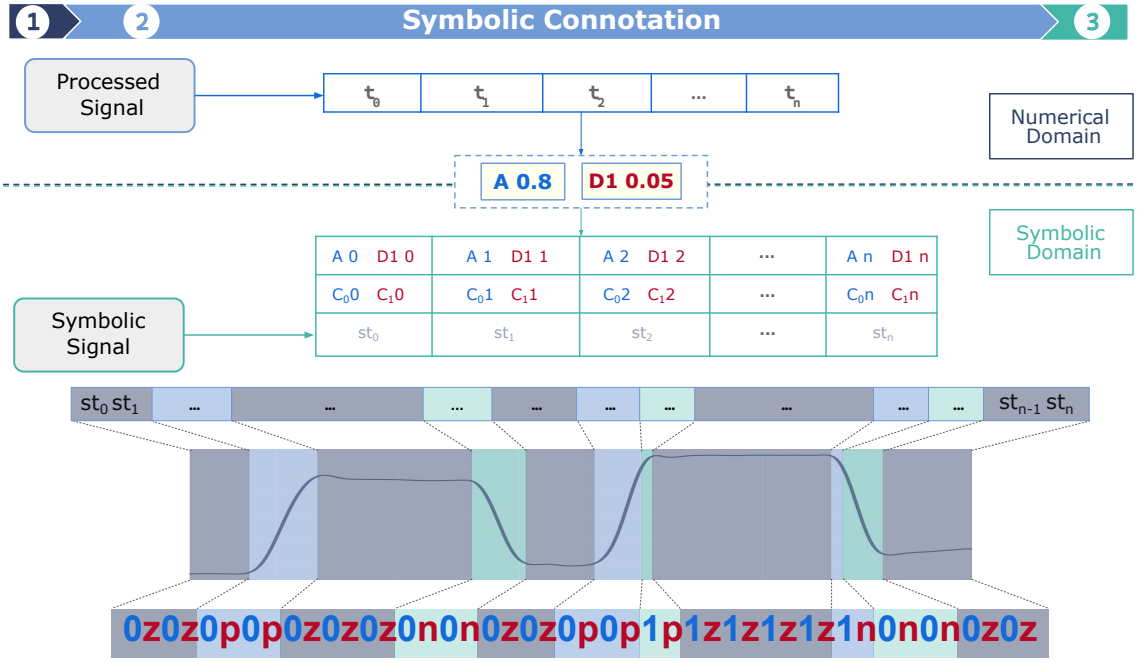


Figure 5.3: Here are highlighted the second stage of the process. It is the moment the signal is transformed from the numerical domain to the symbolic domain. A represents the amplitude method (blue) while $D1$ is the first derivative (red). The exemplified signal plus the symbolic translation are presented.

each sample of the signal into a *character* or group of *characters* that represent the state of the sample for that conversion rule (e.g. if the time series sample has a negative derivative, it will have the character `negative`). Each of the *connotation* methods is related to specific attributes of the time series that are considered relevant for the search procedure. The string that results from this step is therefore a symbolic representation of the strictly necessary attributes of the signal that are best suited to specify a pattern match. Note that this process is decided by the analyst.

The string can be based on a single connotation method or a combination of multiple ones. In addition, the *connotation* process also handles multidimensional time series, in which cases the string is a combination of *connotation* methods corresponding to each time series, returning a group of *characters* for each of its samples.

Following the previous description, we recall the general text definitions from Section 2.7.1 and associate these with *connotation* step's output. Each sample of the time series, $T = (t_1, t_2, ..., t_n)$, is converted into a *character* from this *connotation* alphabet ($char_C$), generating a symbolic time series, $ST = (st_1, st_2, ..., st_i, ..., st_n)$, in which $st_i = char_C$. A *character* is any unit symbol from the connotation alphabet ("Group of Symbols" in Table 5.2). When the analyst selects more than one *connotation* method to transform the time series T, each sample of T is converted into the corresponding *characters*, one for each alphabet of each *connotation* method applied, resulting in: $st_i = < char_{C1} >< char_{C2} >< ... >< char_{Cc} >$, being $char_{Cc}$ the c^{th} *connotation* method applied. In case T is a multidimensional time series, the symbolic time series (MST)' samples from each time series are grouped $MST =$

$(< st_{1,1} >< ... >< st_{1,k} >, ..., < st_{n,1} >< ... >< st_{n,k} >)$.

The set of connotation methods can be defined by the user and added as needed. A base list of connotation methods and the corresponding symbols/tokens used for the examples presented are listed in Table 5.2.

Table 5.2: List of base SSTS connotation operators. The input parameters are s, which represents the input signal and *thr*, which defines the threshold percentage value of the amplitude range of the signal ($max(s) - min(s)$) for a given connotation method. The operator that separates the connotation methods applied to multiple signals or multiple representations of the same signal is the vertical bar "|".

Name	Input Parameters	Group of Symbols	Description
A	(s, *thr*)	[1, 0]	Amplitude comparison, If t_i > *thr**($t_{max} - t_{min}$), t_i is "1", else t_i is "0"
D1	(s, *thr*)	[p, n, z]	Derivative of signal s (s'). If t'_i > *thr*, t'_i is p; elif t'_i < - *thr*, t'_i isn't; else, t'_i is z.
D2	(T, *thr*)	[D, C, z]	Second derivative of signal T (T''). If t''_i > *thr*, t''_i is D; elif t''_i < - *thr*, t''_i is C; else, t''_i is z.
AD	(s, *thr*)	[1, 0]	Determinates if amplitude from a minimum to a maximum and vice-versa is superior to *thr*. If so, t_i is 1, else, t_i is 0.
SA	(s, *thr*)	[r, R, f, F]	Amplitude of a slope segment. The characters are case sensitive, being **r** for a low rise and **R** for a high rise. The same for falling (**F** or **f**).
SS	(s, *thr*)	[r, R, f, F]	Speed of the signal measured by the amplitude on the first derivative. When the sample is quicker than the threshold it is converted to R - quick rise or F - quick fall, whereas when slower it is converted to r - slow rise or f - slow fall.

The symbolic connotation step of the exercise of Figure 5.2 demonstrates the latter formalism. In Figure 5.3, the processed signal is decomposed into a symbolic sequence by two connotation methods: amplitude (blue) and derivative (red). The first implies that the sample values superior to the threshold ***0.8*** will be 1, while the rest turns 0, whereas the second gives the value p when the signal is increasing, n when decreasing and z when stationary with a threshold of ***0.05***. Each of the samples of the signal is converted into the primitive ($st_1 ... st_n$) which is the alternate association of the two connotation methods. The

Figure 5.4: Step 3 of SSTS. Specific *subsequences* of the time series can be searched with a regex. In this case, the search is to find the moment a `plateau` starts to fall, which is found with the regex `z1n`. Blue are the symbols for amplitude and red for the first derivative.

approximate result is illustrated by the bottom string, in which can be seen the alternation property and what it translates. The representation made using these two methods is expected to be necessary to ease the search procedure.

5.1.3 Expressive Syntactic Search

The last step of the process involves searching for the desired pattern in the generated symbolic time series with a regex. This string is governed by the rules of regex from the *alternative regex Python module* [51] and benefits from all the functionalities and meta-characters typically used with this tool, which can be recalled from Chapter 2 (see Section 2.7.4). The search procedure returns the intervals at which positive matches occurred. It is relevant to note that it has been decided to use regex, although any other parser, conveniently combined with the previous steps, might be used for the same purposes.

In the example of Figure 5.2, the purpose is to determine when a plateau superior to *80 %* of the amplitude range of the signal starts to fall. The symbolic time series matches this moment when a the sequence of characters starts with a `1` (*time series is higher than the 80 % threshold*) and is followed by the sequence `zn` (*a sequence of stationary to falling*). The regex used for the search is precisely `z1n`, which indicates that the signal, at some point superior to the threshold will start to decrease, as presented in Figure 5.4.

In the next Chapter we will show several applications of this method for pattern search tasks. For now, we present how to use SSTS for classification purposes.

5.2 HeaRTS - Towards Time Series Classification with SSTS

The ability to transform a time series into a meaningful symbolic representation makes the usage of text-mining tools available for time series analysis. Profiting from such methods can result in different outcomes that can complement the ones resulting from current strategies, improve results or even open novel pathways for further research experiments in this domain. We propose to use the existing text-mining knowledge on this novel representation for time series classification. In order to perform a class separation, differences and similarities have to be highlighted, and this is also possible to be done with text.

In the text mining domain, the differences between text *documents* are traditionally computed with a feature extraction process on text, such as BoW or TF-idf, returning a statistical representation of the words or n-gram (recall Sections 2.7.1 and 2.7.2). *Documents* with different words and/or sequences of words will have different weights on these models, which are used to perform a class separation on the *documents*. We take inspiration from this process and apply it to time series. For instance, the following time series can be very broadly translated into `Flat Rise Fall Rise Flat` or `Flat Peak Valley Flat`, while the following signal would be translated into `Flat Fall Rise Flat` or `Flat Valley Flat`. This type of description is easily performed with SSTS, as showed on Figure 5.7. Figure 5.5 shows the main steps to perform a transformation of the time series into a BoW/TF-idf model that can be used with a classifier.

In this section, we explain each step of the proposed strategy to perform time series classification based on a textual description with SSTS. The fact that the time series is translated into text also provides a layer of readability and interpretability to the data, as

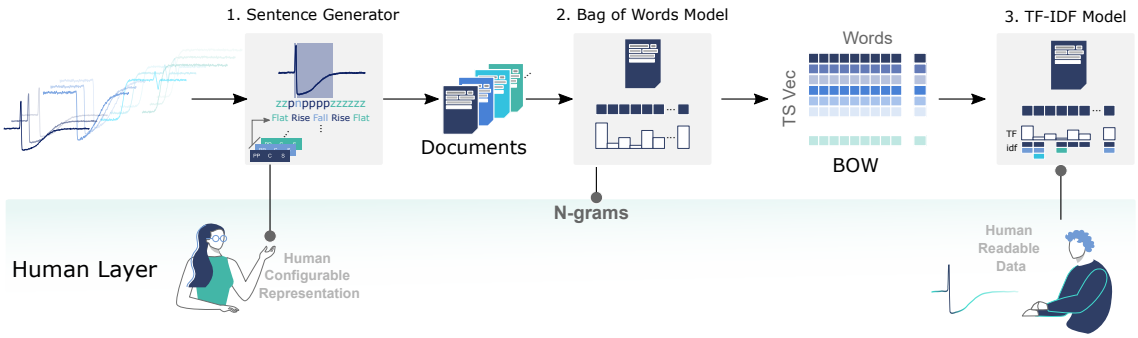


Figure 5.5: Steps of transforming time series into documents and corresponding BoW and TF-idf matrices for posterior classification. The steps are (1) *sentence generation*: all time series are converted to time series *documents* multiple SSTS queries (e.g. `Flat`, `Rise` and `Fall` from the derivative *connotation*); (2) *bag of words model*: each document is converted to a vector with the frequency of words in that document; and (3) *tfidf model*: transforming the BoW to TF-idf and possibly searching for relevant segments of the time series. The human layer shows where the human can apply his/her knowledge, namely selecting the SSTS queries and interpreting the highlights on the signal.

we will explore throughout this section as well.

5.2.1 SSTS to Generate Time Series Documents

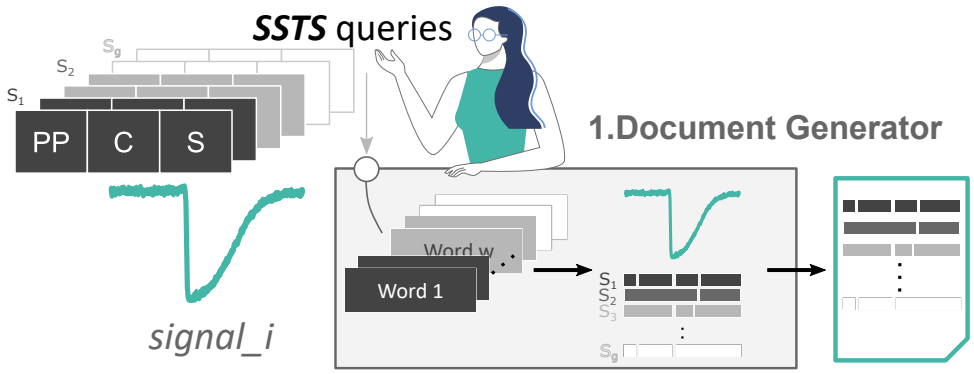


Figure 5.6: Steps for the generation of sentences from a raw time series and organization as a time series *document*. Here: PP - Pre-processing, C - connotation and S - Search are each SSTS queries used to search for the patterns and attribute the matched *subsequences* to a *word*. The colour tones (dark grey, light grey and white) indicate each sentence group ($S_1, S_2, ..., S_g$), which has multiple SSTS queries. The result is a *document* with multiple sentences.

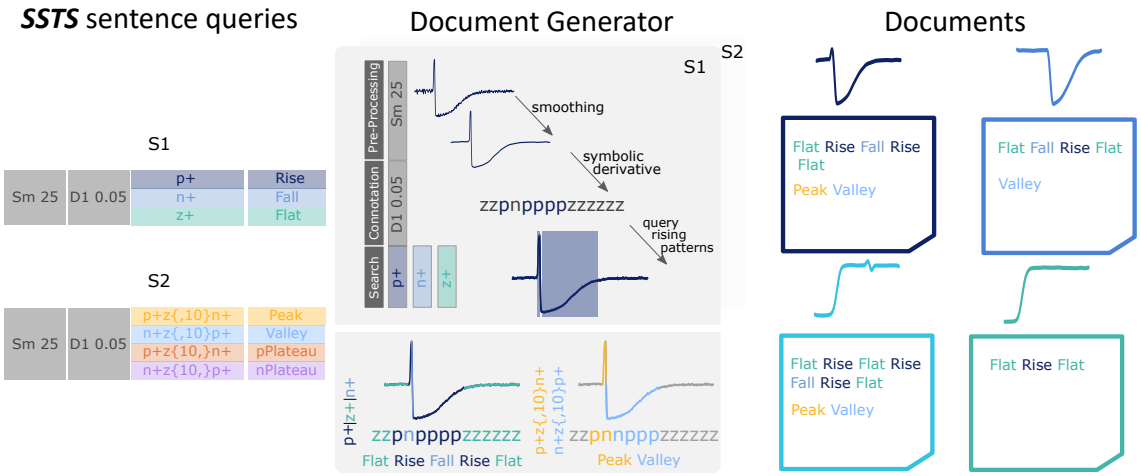


Figure 5.7: Example of converting a time series into a time series document with SSTS. **SSTS sentence queries**: Two groups of sentences (S1 and S2) are used to extract words on a signal. Each sentence group has multiple SSTS queries, to which a word is associated (e.g. `p+` → `Rise`). **Document Generator**.Top: Using SSTS to detect the rising *subsequences* (`p`) of a time series. Each step of the process is written described as follows: (1) *pre-processing*: *Sm* is the function *Smooth* with a window size of 25 samples; (2) *connotation*: *D1*, indicates the first derivate, from which each sample is converted to `z` - Flat, `p` - rising and `n` falling; (3) *search* - regex `p+` searches for all sequences with one or more `p` characters. **Document Generator**.Bottom: Example of sentence generation. Using the other search queries (`p+, n+, z+`), we can find the derivative patterns and convert them into ordered words. **Documents:** the time series documents generated for each of the exemplified signals with the two sentence groups used.

Time series documents are generated by searching for specific patterns and tag the samples of the signal that belong to this pattern with a specific character and word. These are then ordered by sorting the time index they begin with.

In order to perform the pattern search that leads to the signal's description, the words have to be associated with a specific SSTS query. Each query has the three stages of SSTS and searches for the patterns that match on the time series. The patterns that are matched are associated with the corresponding *word*. As an example, the query `{PP: Sm 25, C: D1 0.05, S: p+}` applies a *smoothing* with window size of 25 samples and searches for instances where the time series is increasing. It finally attributes to these *subsequences* the word `Rise`. A specific set of SSTS queries are used to build a full description of the signal dynamics in higher leveled structures, mostly related with amplitude and derivative *connotations*.

The selection of SSTS queries is made by the analyst, either using a pre-defined set of queries or creating his/her own, more appropriate for the type of signals to analyse. This is the customizable step of the process, in which the analyst can include his/her intuition.

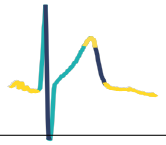
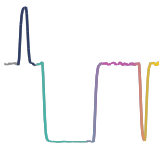
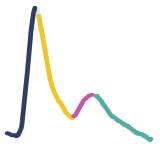
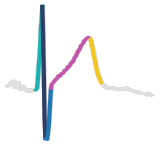
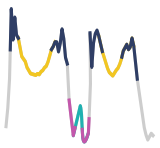
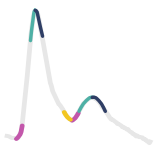
The ability to distinguish time series with text will be as good as the descriptive richness of the generated sentences. Depending on the differences between the time series being classified, a simple description based on the sign of the signal's derivative might be enough (such as simply explaining if the signal is `Rising` or `Falling`). However, in some cases, the overall dynamic of signals might be dissimilar to other dimensions.

In order to have a rich foundation to perform a sentence description of a time series, we created a standard set of SSTS queries, which are presented on Table 5.3, and grouped by sentence ($S1, S2, ...$). The table also shows an example of matching the corresponding query on a real signal. We would like to highlight that these connotation methods and resulting queries are not mandatory and can be adapted by the analyst. All depends on the problem being solved and what might make more sense in this context. In cases where the user knows exactly what kind of patterns are relevant to perform the distinction, specific queries can be used, targeting the relevant differences. Otherwise, the existing list can be used.

An example of performing this analysis on a signal is presented in Figure 5.7. The process highlights the usage of SSTS queries from two different groups of sentences. From the first group, the signal is analysed in terms of derivative, matching parts of the signal where it *rises*, *falls* or is *flat*. A sentence is generated, describing it as `Flat Rise Fall Rise Flat`. The same process is done now with the second group of SSTS queries, but this time, searching for *peaks, valleys, and plateaus*. A sentence is then generated based on the matches: `Peak Valley`. The same process is applied to each class of signals, being, for each signal, generated a *document* with the corresponding sentences.

Having now a method to translate time series into text, we can use text mining methods directly on the *documents*. Typically, text data is vectorized with the BoW or TF-idf models, a process explained in the next Section.

Table 5.3: The connotation variables, search regex and corresponding words assigned to the pattern searched. The parameter *m* indicates the size, in samples, of the difference between a peak or a plateau, *thr* is the threshold for the derivative functions. Each word has a representation on an example of a signal.

Sentence Group	Connotation	Search	Word	Example
S1	D1 thr	p+	Rising	
		n+	Falling	
		z+	Flat	
S2	D1 thr	p+z{,m}n+	Peak	
		n+z{,m}p+	Valley	
		p+z{m,}n+	posPlateau	
		n+z{m,}p+	negPlateau	
S3	SA thr	r+	smallRise	
		R+	highRise	
		f+	smallFall	
		F+	highFall	
S4	SS thr	R+	quickRise	
		r+	slowRise	
		F+	quickFall	
		r+	slowFall	
S4	A 0.5 D1 thr	(0p)+(0z)*(0n)+	bottomPeak	
		(1p)+(1z)*(1n)+	topPeak	
		(0n)+(0z)*(0p)+	bottomValley	
		(1n)+(1z)*(1p)+	topValley	
S5	D2 thr D1 thr	(Dp)+	concaveRising	
		(Dn)+	concaveFalling	
		(Cp)+	convexRising	
		(Cn)+	convexFalling	

5.2.2 Vectorization of Time Series Documents

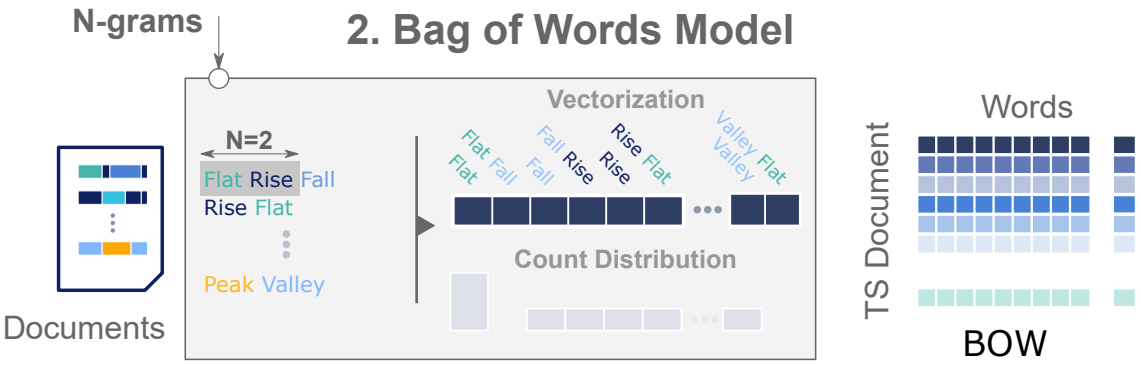


Figure 5.8: Process to transform a time series document into a vectorized representation of words and n-grams. The user can set the size of the n-gram. The documents are transformed into the BoW having a count distribution, which is the tf.

The document generation stage receives a time series and transforms it into a *document* with several sentences, descriptive of its shape and dynamics. As mentioned, this transformation is made with SSTS, which searches for specific patterns on a symbolic representation of the signal and for which a word is given.

After converting the time series into *documents*, the text is analysed to build a matrix of *word* and/or *n-gram* frequencies, the BoW model. In this case, the features extracted are purely statistical, but can provide a relevant measure of differences between documents. Each row of the BoW model is a time series *document*, represented by columns that have the number of occurrences of a *word* or *n-gram*. Figure 5.8 shows the vectorization of a time series *document* as one row of the BoW model.

The BoW model can be used in several ways. Following guidelines from the text mining domain, it can be used for unsupervised or supervised tasks, for topic modeling and keyword extraction. The BoW model can also be trained with naïve Bayes classifiers and linear SVM[150]. In addition, the BoW model feature matrix can be converted into the TF-idf feature matrix, which can then be trained with the same classifiers. In this work, we explored several of these combinations to understand which can give better results.

As was pointed out in Chapter 2 (see Section 2.7.2), one of the reported issues of the BoW model is evaluating the relevance of a word based on its frequency in all documents, while not considering that words that occur in all documents might be less relevant. Therefore, we decided to use the TF-idf model, which increases the relevance of a word using its raw occurrences, while reducing its importance in proportion to the number of *documents* that contain that word. The model is defined by being a ratio between the tf and the inverse document frequency (idf) (equation 2.19).

Both the BoW and TF-idf models give a weight to each *word* for each *document*. This means that each time series *document* is vectorized into a distribution of weights for each *word*, according to its presence on the time series *document* and, in case of the TF-idf model, all the other time series *documents*. The difference between the distribution of each

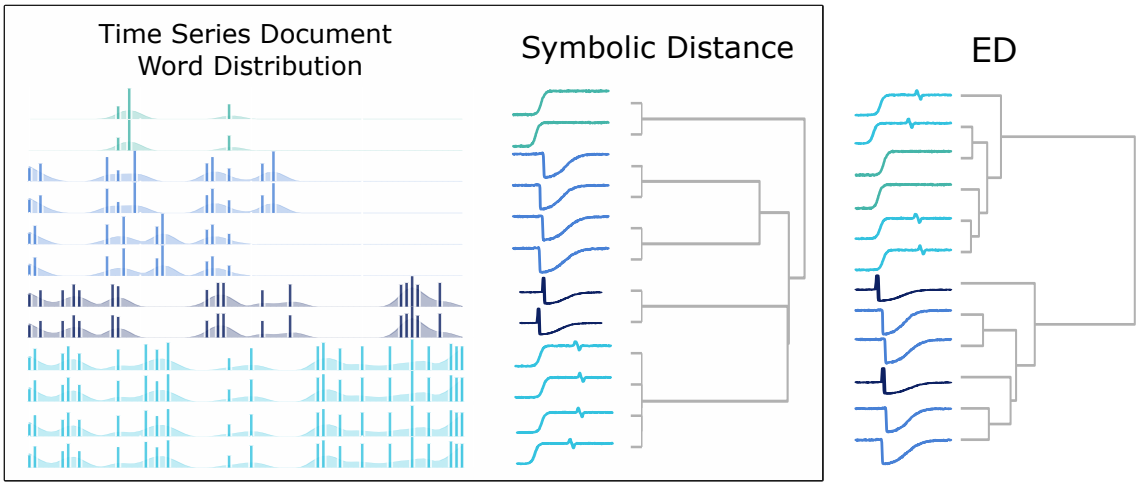


Figure 5.9: Comparing the clustering process of the ED and the proposed symbolic method (Symbolic Distance). In the proposed method, each signal has a word distribution vector, represented on the left. The signals are clustered based on these.

document can be used with a conventional distance measure to compare them. In Figure 5.9 is shown the distribution of *word's* weights for four different time series *documents* of the *Trace* dataset (from Dataset 1, *UCRC*, see Section A.1.1), based on a TF-idf model. The x-axis of each distribution is the *word* or *n-gram* and the y-axis, the corresponding weight for a time series document.

The reader can notice that each distribution is different and this provides a way of distancing time series based on them. On the right of the distribution, the reader can see a dendrogram of sorting examples from the *Trace* dataset. We used the cosine distance between vectors of time series documents (Symbolic Distance) and compared it with the ED. Surprisingly, the ED misses a few matches. Although the example is very simple, it highlights well how this strategy can compensate for some mismatches that can occur with the ED.

As the reader can notice, the ED mismatched signals from class 1 with signals from class 2, and signals from class 4 with signals from class 3. The fact that a misalignment between the main structures of the signals occur (the main peaks from signals of class 1 are misaligned and the main small peak and valley shapes of class 3 are slightly misaligned), increases the distance between signals that have the same shape. On the other end, the proposed strategy follows a distance measure based on the presence/absence of specific shapes, as well as how these are ordered (if *n-grams* are used). This means that classes 1 and 2 will not be mismatched, because signals from class 1 have a peak whereas the ones from class 2 don't.

5.2.3 Towards Interpretable Results

Having demonstrated that it is possible to create a distance measure between time series with pure text, we highlight an advantage of working on the text domain, which is

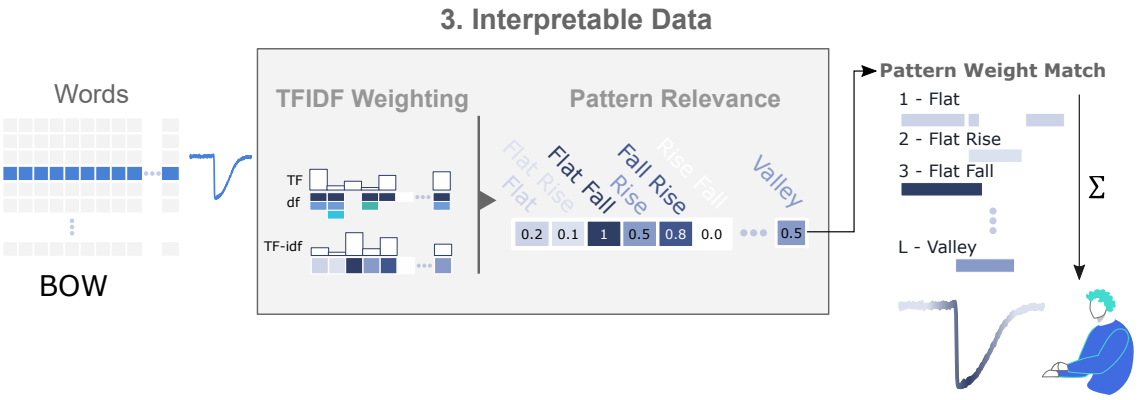


Figure 5.10: From the TF-idf matrix has weights for each word or n-gram (pattern relevance). The search of these words and cumulative attribution of weights on the segments matched can represent a feedback tool. In this case, the valley of the signal is highlighted.

interpretability of the data. By *interpretability*, we mean that it is possible to extract meaning in what differentiates classes of signals. This advantage is taken because of the weighting factor attributed to each *word* or *n-gram* from the TF-idf model. As explained by Pavel Senin *et al.*, the weights can be used as a weighting vector for each word extracted, highlighting the corresponding shape on the original signal and measuring its importance for the classification process [174].

The process to highlight the areas of the signal that are more relevant and able to explain the difference between classes is illustrated in Figure 5.10. From the BoW model, the TF-idf model is computed, following equations 2.17, 2.18 and 2.19. From this representation, each *time series* is represented as a vector of weights for each *word*. These weights can be used to highlight the area of the signal represented by the corresponding *word*. Therefore, the process is to search back the areas of the signal that match the *word* or *n-gram* (w_i) as an SSTS query, while keeping the corresponding *weight* (h_{ij}), for word i from time series document j, on the indexes of the match, with indexes a and b:

$$(a, b) = ssts(t_j, w_i) \tag{5.1}$$

In the end, the weighted vectors are summed, returning a final vector (WTS) with the weighted contributions of all *words*:

$$WTS_i(a, b) = WTS_i(a, b) + h_{ji} \tag{5.2}$$

In Figure 5.10, the exemplified signal has a higher relevance on the `valley`, being the most characteristic element of that time series.

5.3 Towards Natural Language for Pattern Search

The fact that an analyst can search with regex on time series makes the process more flexible and expressive and is innovative in the sense that text patterns can be written

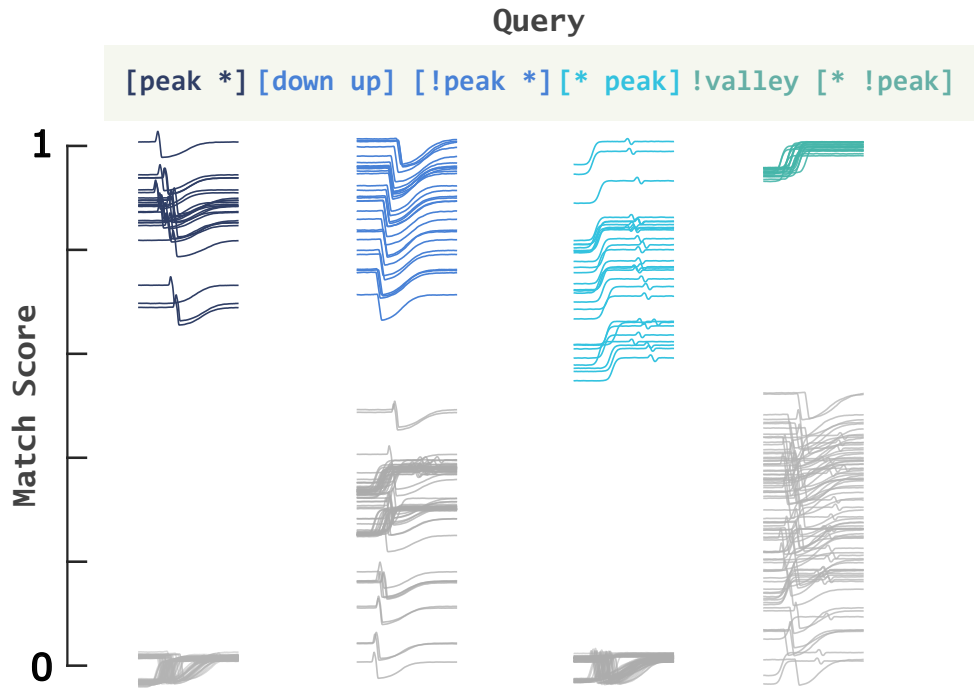


Figure 5.11: Using QuoTS, we can create short, intuitive natural language queries to rank the 100 exemplars in Trace, separating our sought class from the remaining data. Here $\star$ is *anything*, ! is *not*, and *square brackets* are a grouping operator (more details in Section 3).

to search for specific shapes on a time series. In this journey to explore the topic of text-based queries for time series pattern search, we established the possibility of making an association between features and keywords to perform a search similar to what is done with web-searches toolbars. Thus, we propose a novel method for pattern search, QuoTS. This method is based on natural language keywords and linguistic operators. In contrast with the previously explained method, SSTS, QuoTS associates *words* with features, and such as SSTS, contributes to explore time series with more expressiveness and ultimately to democratize it.

5.3.1 Browsing Patterns Search

When a user opens any search toolbar (such as *Google* or *Safari*) and searches for a specific *topic*, he/she will write a set of keywords that can match the *topic* he/she is searching. The results of this process are a list of pages ranked based on how well they matched the *keywords* used. Of course that in this example, search toolbars might have additional information (maybe including geolocation, browser history, etc...), and not solely rely on the *keywords* for the search process. However, let us take into account this simple *keyword* based search process that we typically use when browsing the internet and apply it to pattern search on time series.

As we mentioned above, patterns can be described with *words* by associating text with specific properties or the presence of specific shapes on the pattern. In that sense, we propose to create a search paradigm that relies on words represented by features extracted

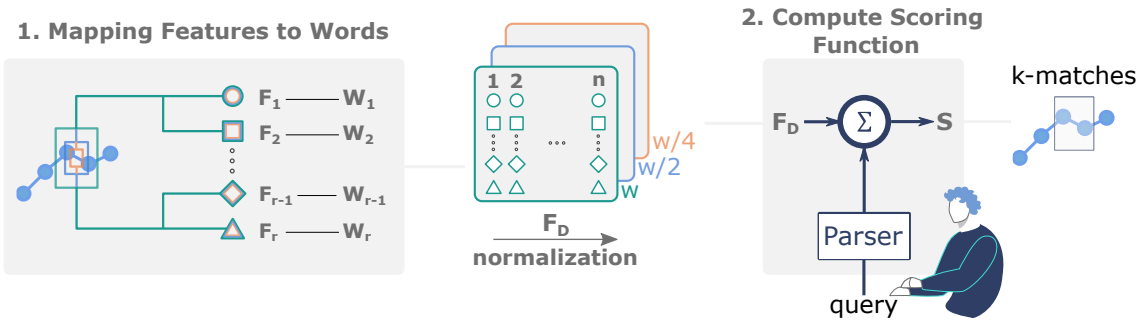


Figure 5.12: QuoTS steps. It shows the feature extraction process and matches each feature to words ($W_1, W_2, etc...$). These features are computed in three window (w) dimensions (w, $w/2$ *and* $w/4$). The user can input a query to search for a specific pattern. This query will be scored based on the parser and features. It then returns the top k-matches.

from the signal. Informally, if a user wishes to find examples of a particular behaviour, he/she can simply *Google it*, by describing the data he/she expects to see, given that behaviour. For example, we considered the four-class *Trace* dataset (from Dataset 1, *UCRC*, see Section **??**), which has been studied in more than 1,000 papers. As Figure 5.11 shows, it is possible to use our system to create queries that can separate each of the four classes from the rest of the data. This separation is possible simply by describing with *words* the presence/absence of structures. Further, we will explain in detail how the process is done, which are the operators that we designed, and how queries can be written. Still, for now, we will explain the meaning of each query in this example. For instance, signals from class 1 () match well with query `[peak *]`, which translates into *has a peak on the first half and anything on the second half*. The method sorts the signals based on this query and gives a very low score for all signals that do not have a peak in the first half of the signal. The same happens for the other queries, used to sort the other classes. A similar result is achieved with the inverse query `[* peak`, which means *anything on the first half and a peak on the second half*, as well as the query `!valley [* !peak]`, which translates into *has not a valley and has no peak on the second half*.

The overall steps of the process are illustrated in Figure 5.12. In order to perform the search with natural language keywords and operators, we represent the time series into a set of feature series. The features are extracted with a *moving window*, with total overlap. The process is made three times, with decreasing window sizes (w) dimensions ($w, \frac{w}{2} and \frac{w}{4}$). Each feature is associated with a *word* (W) that should give a good intuition about the property it represents (e.g. `up` or `rising` should be clear to be indicative that the signal is rising). This feature set is the weight of the *word* for each *subsequence* of the signal. Such that when the *words* are written, the features are summed according to the operators used. Then, the search returns the *k-top subsequence* matches for the query used.

Concretely, with this method, we will demonstrate that the proposed natural language keyword representation is expressive enough to discriminate specific behaviours on a time series. Let us dwell into how to map features to words.

5.3.2 Mapping Features to Words

We define a *word feature vector* W as a mapping of a feature series F_i with a specific *word* W_i. With feature, we intend to describe either a property, such as the mean or standard deviation, but also a distance measure to pre-defined examples. In linguistic terms, this *word feature vector* is an adjective of the *subsequence*. Every *subsequence* $t_{i,m}$ from each time series T is characterized by the selected set of *words*, being created a set of *word feature vectors* for each time series, further normalized between 0 and 1. The *word feature vector* has the size of T and the feature series values indicate how relevant is the *word* in the *subsequence*. Figure 3 shows an example of the word feature vectors for up (Fup) and down (Fdown).

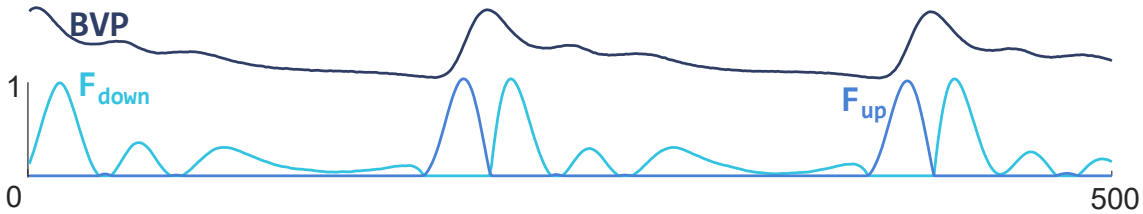


Figure 5.13: Example of a word feature vector. In this case F_{down} and F_{up} represent a negative and positive slope value.

To allow interactive search, the *word feature vectors* are pre-computed in an offline indexing stage. In addition to this, we also extract three dimensions of the same word feature vector with different window lengths, based on w: $W_1 \rightarrow w$; $W_2 \rightarrow \frac{w}{2}$; $W_3 \rightarrow \frac{w}{4}$; with the intent of matching ordered sequences of words inside a subsequence with the *grouped followed by* operator, as will be explained further. It is important to note that the ratio of the window used for the dimension W_2 and W_3 might have to be fine-tuned for the domain, as there might not be a "one window ratio fits all". However, empirically we observed that the exact value of this ratio is not critical to the success of the search.

For each W is assigned a *word*. We use *English* words to make the process more intuitive, such as `noise`, `up` and `peak`. We recognize that the intuitive meaning of such words can vary from user to user depending on their domain, their experience, and the current context. Either way, words can be mapped to features that are domain-specific or word feature vectors can be given a domain-specific vocable, providing a more appropriate mathematical thinking behind what is its meaning. In addition, we are aware that multiple words can be given the same meaning and for this reason, we associate several synonyms with each word.

We define an initial subset of features that are mapped to words. When defining one *word feature vector* it would often come with a negation pair. A negation pair (!W) represents the exact opposite of a defined *word feature vector* (W) following the rule:

$$!W = 1 - W \tag{5.3}$$

This indicates that when one increases the other one has to decrease proportionally. Examples of such *word feature vectors* are `symmetric` and `asymmetric`, or `complex` and `simple`. Note that some words might be the opposite of each other, but do not follow this rule, or even seem to be the opposite of each other, but are not. For instance, `up` and `down` are opposite of each other, but do not follow Equation 5.3. While one exists, the other cannot, but it does not mean that when one is small, the other has to be high since the *subsequence* might just be `flat`. Another case is the word `peak`. Intuitively, we would think that `valley` is the opposite of `peak`, but the consideration of `!peak = valley` (*not being peak means it is a valley*) is false. In this work, we use the negation of a word feature vector for cases where there is no negation pair. This negation is realized using an operator (See Table 5.4).

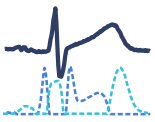
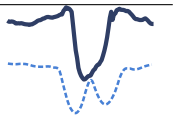
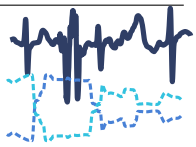
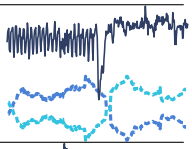
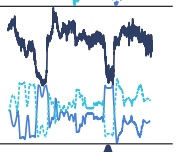
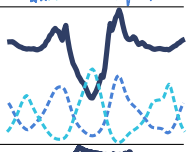
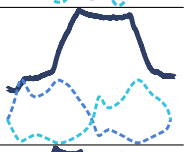
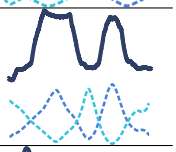
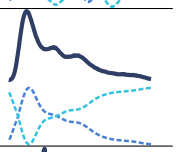
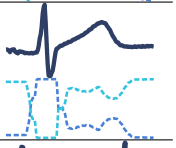
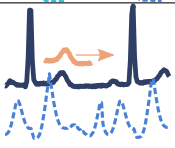
As previously noted, a set of features is used to extract several properties of all *subsequences* of a time series and attribute a semantic meaning to each one of them by mapping it to a specific *word*. It is our assumption that a *subsequence* can be mapped to a set of *words* that an analyst would use to describe it. Depending on the domain or *vocabulary* of the analyst, the set of words might have to be different and adjusted. Eventually, the *dictionary* can be expanded to other types of *features* and *words*. In any case, we want to demonstrate that this current set of words and operators can solve many search problems with expressive queries.

5.3.3 Linguistic Operators

In the same way we use word and sentence connectors in our language to create contrast or attribute a temporal sequence, in our proposed system we use *operators*. An operator is a metacharacter or a word that can be used to diversify the way word feature vectors are handled, either in the way the information is extracted or how these are combined. It contributes to a more versatile and expressive usage of this language. Currently, we have a simple list of four operators: *negation* (`!`), *wild card* (`*`), `followed by`, and `grouped followed by` (e.g., $[W_1 W_2 \dots W_w]$. This list can be expanded and customized, but we want to demonstrate that with a minimal set of operators, most of the problems we present are solvable.

Web search engines have many operators at the user's disposal, but since a list of words is usually powerful enough to retrieve and correctly sort most of the desired results, very few (or none!) are often used. We believe that this is the case for this application as well but acknowledge that simple operators can make the query more natural and come in handy to perform conjunctions between features and multiple dimensions, such as *temporal logic* or *negation*. For instance, we often rely on *temporal logic* to express the presence of a shape in regards to another: `the peak that comes after the valley`, or even use the absence of a property: `it does not have a peak`. We also typically express the order in which structures occur on a time series: `up and then down`. These operators are especially useful to close the gap between the query and human discourse, contributing

Table 5.4: List of all word feature vectors with examples. Here, $i \in [0, n]$, n is the signal's size, a is the beginning of the moving window, starting at $a = i - \frac{w}{2}$, and w is the moving window size. The colors of the *words* are the corresponding colors on the *example*. Each example has representative feature vectors. When a negation pair is present, it is added to the example.

Word	Description	Example
Up (**Down**)	The slope estimation of a linear adjustment ($y = ax+b$) to the *subsequence*, being `up (down)` $= a$, if $a_{\texttt{up}}$>0 ($a_{\texttt{down}}$ <0) or `up (down)` $= 0$, if $a_{\texttt{up}}$<=0 ($a_{\texttt{down}}$ >=0)	
Flat	The inverse of the sum of absolute differences between the *subsequence* and its average: `!flat(i)` $= \Sigma_{j=a}^{a+w} \lvert t(j) - \overline{t(a, a+w)} \rvert$.	
Noise (**Smooth**)	The residual error when modeled by a moving average (ma). `noise(i)` $= \Sigma_{j=a}^{a+w} \lvert t_j - ma_j \rvert$. (`smooth = !noise`).	
Complex (**Simple**)	The complexity-invariant distance measure of the *subsequence* of 2.14. (`simple = !complex`)	
Symmetric (**Asymmetric**)	The normalized ED (2.8) to the *subsequence*'s horizontally flipped self.	
Peak (**Valley**)	The logarithmic normalized ED (2.8) to the template of a peak (valley), modulated by a gaussian function.	
StepUp (**StepDown**)	The logarithmic normalized ED (2.8) to the template of a step-up(down) function.	
PlateauUp (**PlateauDown**)	The logarithmic normalized ED (2.8)to the template of a plateau-up(down) function	
Top (**Bottom**)	The moving average of the time series: $ma(i) = \frac{1}{w}\Sigma_{j=a}^{a+w} t(j)$. (`bottom = !top`)	
High (**Low**)	The difference between the maximum and minimum value of a *subsequence*: `high(i)` $= max(t(a, a+w)) - min(t(a, a+w))$. (`low = !high`)	
Shape	The normalized ED profile of the time series with a query given by the user as an example. A word must be given as well and integrated into the query language.	

to a more expressive mechanism when using the proposed language. Currently, four operators are available. Below is a list and description of each of them, starting with the negation operator.

- **Negation Operator - `!W`** : As mentioned above, most words come with an opposite pair, but some do not follow Equation 5.3. In these cases, or when the word has no direct opposite, it can be useful to penalize the presence of the word in a *subsequence*. This operator does that by applying Equation 5.3 to the word feature vector, W.

  When describing time series, we inevitably use temporal logic in explaining the sequence of shapes we perceive. The next operator is `followed by`.

- **`A followed by B`**: This operator rewards a *subsequence* represented by A followed by one *subsequence* that has a high score for B, within a distance of size w. A and B can be single words, multiple words, or even queries for different dimensions of the time series.

  With this operator, we look ahead of a *subsequence* in the time series. However, in some cases, it might be useful to describe the sequence inside the limits of the window we defined. For these, we have a special case of `followed by`, which is the `grouped followed by`, represented by square brackets ([ ]).

- **Grouped `followed by` -** $[W_1 W_2 \ldots W_N]$: Instead of looking ahead in the time series, we look inside the *subsequence* to reward an ordered sequence of words. In this special case, the *subsequence* is segmented into N sub-windows, with size integer of $\frac{N}{m}$, and the corresponding *word* is scored within this sub-window. For this, we use the other two dimensions of the word feature vectors (W_2 and W_3). If $N <= 3$, W_2 is used, while if $N > 3$, W_3 is used.

  Often, within a *subsequence*, the differentiating property occurs on the first half, last third or another sub-window of the *subsequence*, while the remaining sub-window(s) is(are) not relevant. We, therefore, introduce the wildcard ($\star$) operator.

- **Wildcard - *:** The sub-window where * is used is valued equally for all *subsequences*.

As with vocabulary, the reader could imagine expanding our dictionary of operators, but even with a limited set of them, we can successfully solve all the proposed search tasks, which cover dozens of examples. After presenting the set of elements that can be used in the proposed method to query a pattern of interest, we are ready to explain how the query is turned into a score function and finally how the selection of the k-most relevant *subsequences* is done.

5.3.4 Natural Language Query for Time Series

As illustrated on Figure 5.12, the user inputs a query with all the *words* and *operators* available. The query is parsed to calculate a matching score (S) that indicates which

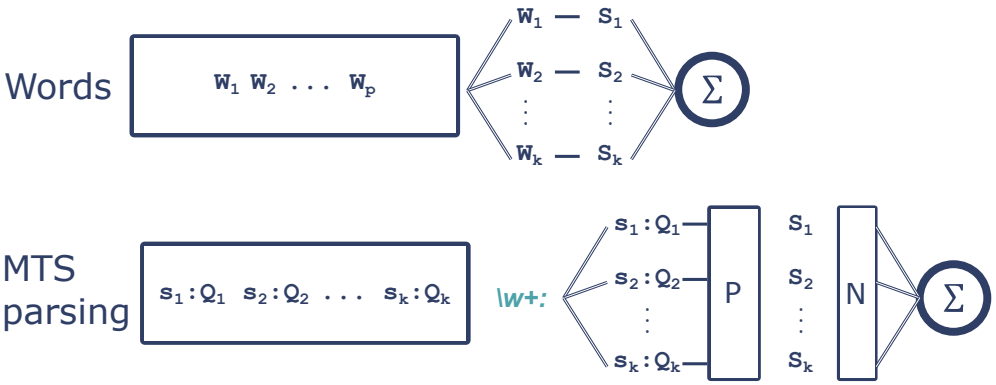


Figure 5.14: Process to parse the input text query. When having simple space separated words, each corresponding word feature vector is summed together. When in multidimensional mode, there is the *signal id* first.

subsequences are better represented by the query. The score is calculated based on the word features vectors extracted from the time series that are *called* by the query.

As mentioned, considering the possibility of using the `grouped followed by` operator, each feature is extracted three times with different window sizes. For each word, three sets of word feature vectors are stored ($W_1, W_2 and W_3$) based on the original window size ($w, \frac{w}{2} and \frac{w}{3}$). These word feature vectors are stored together in a *dataframe* and called based on the *word* written by the user. It is important to mention that all words are stored in a vocabulary file, associated with a *thesaurus* file for synonym checks.

If the signal is multidimensional, all three sets of *word feature vectors* are extracted for each dimension. Having this set of information pre-computed helps make the search run at interactive speeds, even for large data collections. When all data is pre-computed, QuoTS is ready to accept queries by the user.

The query field accepts any word available in the vocabulary and *thesaurus*. When any of the words are not present in our vocabulary, we alert the user which is the closest word available, based on *edit distance*. The query can accept operators and works for multidimensional querying. These are relevant elements that are used as a reference to parse the query into individual scoring elements. This parsing process is made by looking into: (1) which dimension(s) of the time series is (are) included; (2) which operators are used; and (3) the single written words. The first two define how the score is calculated, that is, how word feature vectors are combined, as well as which dimensions of the word feature vectors are used. For instance, when including multiple signals, the query parses which word(s) corresponds to which signal, to search for the correct index of word feature vectors in the pre-computed *dataframe*. To identify the signal, the following regex is used to find all signal names *\w+:* (*one or more characters ending with ":"*).

When the `followed by` operator is used, the query is parsed in the *elements* that come before and after it (this is applicable either if the operator is used for intra-signal or inter-signal search). For this, the following regex is used to separate the *words* before and after the operator: *\w+ followed by \w+* ((word or words before followed by and word

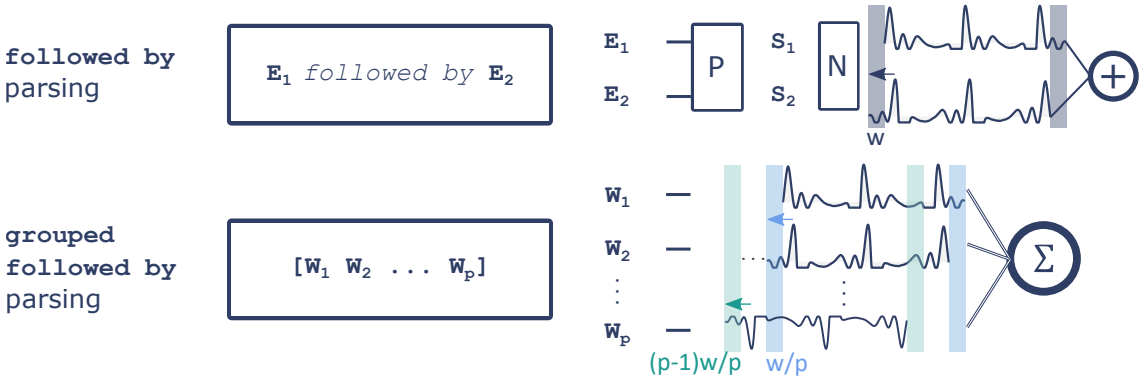


Figure 5.15: Second order of the parser. In case the `followed by` operator is used, the elements (E_1 and E_2) are parsed (P) and word feature vectors are summed, but the E_1 has to be delayed to count as the "previous" instance (w represents the selected window size). When using the more specific `grouped followed by` operator, square brackets are used and represent the window size highlighting the signal. Only words can be used in these brackets and are added together, delaying each word based on their position inside the brackets, for instance, word 2 (W_2) is delayed by the window size, divided by the number of words inside the brackets (w/p).

or words after it). What is meant by *elements* is either *words* (intra-signal) or two different signal dimensions (inter-signal). Another temporal operator is the `grouped followed by`, which is parsed by identifying square brackets in the query with the following regex `\[.+?\]` (*anything inside square brackets*). The content of the square brackets is then combined (see Figure 5.15).

When any of these elements are parsed, we end up with a single word or sequence of words, which are combined by summing their corresponding *word feature vectors* (this corresponds to an implicit OR). Figures 5.14 and 5.15 give a visual aid to understand the parsing process for each case.

The simplest case for a query would be a sequence of *words*. In this case, each *word feature vector*, associated with the corresponding *word*, is a score (S) function and is summed together to give the final score. In case a multidimensional signal is used, the user can write a multidimensional query, identifying a specific query (Q) for each dimension ($s_1, s_2, ...s_k$). For this, each query Q corresponding to a signal s is treated individually by the standard parser (P on Figure 5.14), resulting in a scoring function (S) for each signal used on the query. Before summing all scoring functions, these have to be normalized (N on Figure 5.14) to have equal weight. It is important to note that Q can be a query with *words* or the `grouped followed by` operator.

The temporal operators are parsed by the regex, as mentioned above. From the `grouped followed by` operator, only *words* can be used and are combined with a temporal delay that depends on the number of *words* inside the brackets. The brackets represent the limits of the *window* that is used to extract the *word feature vectors*. In that sense, when using two *words* inside the brackets, it is expected that the first half of the signal has the first word, and the second half has the second word. This is extended to

additional words inside the brackets. In order to combine the *word feature vectors* of the *words* used inside the brackets, we delay each *word feature vector* based on the number of *words* and corresponding position inside the brackets, such that *words* after the first one are delayed $\frac{(i-1)w}{p}$ samples, being i the position inside the brackets, p the total number of *words* inside the brackets and w the window size. The delayed vectors are added together. For example, if using the query `[up down]` and features were extracted by a window with size of 100 samples, the *word feature vector* for the word `down` is delayed 50 samples and added to the `up` *word feature vector*. When used in combination with individual *words*, the scoring function resulting from this operator is normalized between [0-1]. The reasoning is that each segment of the query should be weighted the same (e.g., if using the query `noise [up down]`, as `up` and `down` are combined, the range of this segment is [0-2], while `noise` is [0-1]. Therefore, `[up down]` is normalized between [0-1] before being added to `noise`). A similar process is performed with the `followed by` operator. The scoring function for the query previous to the operator is added to the scoring function of the query next to the operator but delayed by one window size (w).

Finally, depending on the *words* and operators applied, the resulting scoring function is used to search for the k-top *subsequences*. This is made by searching in k-iterations the maximum value of the scoring function. During this iteration process, trivial matches are not considered. A trivial match is a high-valued sample of the scoring function that is a neighbor of the maximum value found. To avoid these matches, all samples of the scoring function around a k-top match on the scoring function ($[argmax(S) - \frac{w}{2}, argmax(S) + \frac{w}{2}]$) are converted to 0. The matched *subsequences* are highlighted on the signal and sorted by their match score.

6

Experimental Analysis, Validation and Discussion

The previous chapters (Chapters 5 and 4) introduced the methods developed in this thesis. For each topic, a few examples were presented to help understand and follow the explanation and description of the proposed methods. In this chapter, we go further into testing and validating the algorithms to provide evidence that these are relevant and contribute to the state-of-the-art in the time series data mining field. The performance was evaluated on public datasets, introduced in Chapter **??**.

Considering the diversity of methods developed, this chapter will be divided into each contribution following the topics presented in the previous Chapters:

- **Information Retrieval with the SSM**: We presented three main applications of retrieving information from the SSM, namely for novelty segmentation, periodic segmentation and similarity association. The process for novelty segmentation was validated in public datasets and its performance compared with two referenced methods (WL - window based segmentation, and BS - binary segmentation). In addition, we applied this strategy in a change point detection benchmark, where we were able to compare its performance with all methods available in the benchmark. We show an application example for periodic segmentation and query-based segmentation in one public dataset, and results are computed based on ground-truth annotations.

- **Pattern Search with SSTS**: In this topic, we provide use-cases where SSTS is used to retrieve specific patterns. Each step of the proposed method are highlighted. Specific metrics (see Section 2.8.3) are used to compare the SSTS query used with a standard python-based implementation for the search problem.

- **Time Series Classification with HeaRTS**: Having demonstrated the possibility of using SSTS to search for patterns, we explore its usage to extract a textual representation of time series and apply text-mining methods to perform the classification. A public benchmark was used to perform this analysis and HeaRTS was compared

with the 1-NN ED. We also provide evidence in several use-cases of how the weights of the TF-idf model can be used to highlight the most relevant areas of the signal, contributing to the interpretability of results.

- **Pattern Search with QuoTS**: The second proposed method for pattern search in time series will be evaluated with several use-cases. We also evaluate its ability to separate classes of hand gestures with simple queries. Finally, we also show an example of *pupetteering* as a toy problem, and as a possible application of this method.

6.1 Information Retrieval with the Self-Similarity Matrix

6.1.1 Novelty Segmentation

In this section, we present several examples of how the novelty function (recall Section 4.3.1, retrieved from the SSM, is useful for the segmentation of time series. We also provide a measure of the algorithm's performance considering ground truth events from public benchmarks, while also comparing our proposed solution with existing methods for change point detection from the *Turing Change Point Detection Benchmark* [26].

6.1.1.1 Quantiative Evaluation Metrics

In order to evaluate the performance of our proposed method in general scenarios, we applied the algorithm to Datasets 3, 4, 5, 6, 7 and 12 (*HAR1*, *WISDM*, *EMG*, *ECG1*, *ECG2*, and *ENABL3S*, see Sections A.1.3, A.1.4,A.1.5,A.1.6, A.1.7 and A.1.12). The evaluation was divided into *biosignals-related applications*, and the general change point detection benchmark (Dataset 9 - *ATCPD*, see Section A.1.9). The biosignals experiments are associated with public datasets from Physionet, the *UCI Machine Learning Repository* and the *UCR Semantic Segmentation Benchmark*, involving different contexts (HAR, hand posture, and noise detection) and sensor types (ACC, GYR, EMG, ABP and ECG).

The quantitative evaluation on biosignals' public datasets was made by accumulating true positive (TP), false positive (FP), and false negative (FN) values with a tolerance zone around the ground truth events. The applied reasonable tolerance was the ground truth wrapped by a window size of the SSM computation, inside which a detected event was counted as a TP. The case that no estimated event was found inside the tolerance band was considered an FN. An estimated event outside the tolerance or duplicating to an already counted TP were regarded as an FP. The F1-score based on the precision and recall values was calculated from TP, FN, and FP values (recall definitions and equations from Section 2.8.2).

An online repository [1] is also available where Tables with detailed parameter usage

[1] https://github.com/JmdRodrigues/PhDThesis/tree/main/ProjectCode/NOVA/Results/Tables

Table 6.1: Overall results for the performance of the method on novelty segmentation, including true positives (TP), false positives (FP) and false negatives (FN), precision (P), recall (R) and F1-score (F1) values. The last row provides the macro averaged F1-scores (*M.A. F1*) of the four datasets.

	Novelty						WL						BS					
Dataset	TP	FP	FN	P	R	F1	TP	FP	FN	P	R	F1	TP	FP	FN	P	R	F1
HAR1	166	16	13	0.91	0.93	0.92	169	9	10	0.95	0.94	0.95	125	58	54	0.68	0.70	0.69
WISDM	784	87.0	92	0.90	0.89	0.90	709	140.0	167	0.84	0.81	0.82	657	255.0	219	0.72	0.75	0.73
ECG1	18	1.0	0	0.95	1.00	0.97	13	5.0	5	0.72	0.72	0.72	16	2.0	2	0.89	0.89	0.89
ECG2	155	9.0	13	0.95	0.92	0.93	139	13.0	29	0.91	0.83	0.87	122	58.0	46	0.68	0.73	0.70
EMG	695	68	33	0.91	0.95	0.93	608	123	120	0.83	0.84	0.83	351	413	377	0.46	0.48	0.47
Enabl3s	3433	469.0	366	0.88	0.90	0.89	2941	327.0	858	0.90	0.77	0.83	2618	1322.0	1181	0.66	0.69	0.68
M.A. F1	-	-	-	-	-	0.91	-	-	-	-	-	0.83	-	-	-	-	-	0.69

are presented, namely the window sizes, kernel sizes, and thresholds. In addition, the distribution of F1-scores is available in Appendix B Section B.2. We also encourage to use a live version of the method on your own data, which can be found at [2]

6.1.1.2 Biosignals' Segmentation

In the evaluation of biosignal segmentation, the performance of our proposed method was compared with existing approaches available on the Python library *ruptures*, namely the window-based segmentation (WL) and the binary segmentation (BS) [187]. For these methods, we used the *normal* cost function, which accounts for mean and variance changes. The benchmark evaluation referred to the best score obtained from the state-of-the-art methods available on repository [26]. The evaluation procedure to detect TP, FP and FN was the one followed on [26].

The method has been computed in the same conditions and followed the same procedure for all datasets' records. The features used were the same for each record (see Appendix 2 - B), varying the timescale parameter, the overlap size of the sliding window, and the kernel size. The peak detection strategy based on a threshold mechanism is the same for all records, while the threshold value varies from record to record. Results for publicly available datasets are listed in Tables 6.1 and 6.2, and Table 6.3 expands the performance by F1-scores in detecting the change point events from a benchmark.

The illustrative examples provided in Section 4.4 corroborate the proposed method's capability in segmenting real, complex, and multimodal biosignals datasets. As Tables 6.1 and 6.2 convey, an overall macro-averaged 0.91 F1-score is achieved, while the competitors' overall F1-scores are 0.83 (WL), and 0.69 (BS), respectively, with the proposed method winning in 5 of the 6 datasets in the overall F1-score ranking.

For Datasets 3, 4, 5 and 12 (HAR1, WISDM, EMG and Enabl3s, see Sections A.1.3, A.1.4, A.1.5 and A.1.12), the window-based methods, *novelty* and WL, performed much better than the BS method, mainly because the sliding window algorithm with a full set of features comprehensively characterizes changes in the signal. The standard WL uses cost

[2] https://jmdrodrigues-irbiosignals-app-streamlit-main-b7t5rc.streamlit.app/

Table 6.2: Results to summarize the comparison between the methods tested. In this table, each record of each dataset is counted for each method as a win, draw or loss. The count is made for each dataset and finally, for all datasets altogether (W | D | L).

	Novelty			WL			BS		
Dataset	Wins	Draws	Loses	Wins	Draws	Loses	Wins	Draws	Loses
HAR1	0	0	2	2	0	0	0	0	2
WISDM	42	1	8	5	1	45	3	0	48
ECG1	2	7	0	0	4	5	0	6	3
ECG2	10	1	1	0	0	12	1	1	10
EMG	31	2	3	3	2	31	0	0	36
Enabl3s	320	52	104	100	47	329	3	7	466
Overall W | D | L	5	0	1	1	0	5	0	0	6

functions searching for mean/variance value changes in the signal, which achieves a high F1-score in most of the datasets explored, even identifying transitions between dynamic activities, such as *Walking/Upstairs*. Our proposed method had a similar performance with a worse count in FP values, where the added features did not improve the segmentation performance. In contrast, our proposed method, complemented by additional features, had a much better performance than the WL method in Datasets 4, 5 and 6 (WISDM, EMG and Enabl3s, see Sections A.1.4, A.1.5 and A.1.12), having a higher F1-score in a total of 35(/45), 31(/34) and 320(/423) records for each corresponding dataset. Adding features enabled a more robust and sensitive detection of pattern changes, although it missed some changes between similar patterns, like *Walking* and *Upstairs/Downstairs* in Dataset 3 (HAR1, see Section A.1.3), which are the primary source of the FN value. Similar to the FN values, the FP values of our proposed method are mostly superior to other methods. It nonetheless leaves room for discussion. Some events are not marked as changes in specific activities, but the signal pattern actually changes. For example, Figure 4.8(bottom) exposes pattern changes during an *Upstairs/Downstairs* activity unlabelled in the ground truth, possibly due to a flight of stairs. The novelty function is sensitive to such pattern changes, which inevitably contributes to the FP values during the comparison with the ground truth. The higher count in FP might also be caused by the peak search strategy. Being based on a simple threshold, it might detect more peaks that are not relevant, which makes the P comparable to the WL method. Another strategy should be considered for this step.

For Dataset 6 and 7 (ECG2 and ECG1, see Sections A.1.6 and A.1.7) the proposed method was also superior. Although ECG-based jump artefact detection is fundamental, the WL method could not find all the segmentation borders, while the BS method worked better. In Dataset 6 (ECG2, see Section A.1.6), the same ECG signal was noised with different Signal-to-Noise-Ratio (SNR) levels to form a new set of resultant signals. Overall, our proposed method was able to detect the changes between noisy segments and clean segments with noise down to 12 dB. At 6 dB, the proposed method achieved an F1-score

Table 6.3: Comparison of the F1-scores between our proposed method (*novelty*) and other algorithms' benchmarks in Dataset 9 (see Section A.1.9). The calculation of all one-dimensional signals' average performance and all signal's average performance does not include columns with a gray background where no change point should be detected, or a signal error was present. Bold values represent the best F1-score for this specific dataset. T: timed out; F: failed compiling.

Dataset	NOVELTY	AMOC	BINSEG	BOCPD	BOCPDMS	CPNP	ECP	KCPA	PELT	PROPHET	RBOCPDMS	RFPOP	SEGNEIGH	WBS	ZERO
One-dimensional															
bank	0	**1.000**	**1.000**	**1.000**	0.500	0.054	0.200	0.333	0.400	**1.000**	T	0.015	**1.000**	0.043	**1.000**
bitcoin	0.694	0.507	0.690	0.733	0.533	0.611	0.625	0.665	**0.735**	0.446	T	0.284	**0.735**	0.690	0.450
brent_spot	**0.861**	0.465	0.670	0.609	0.239	0.607	0.636	0.553	0.586	0.249	T	0.521	0.586	0.564	0.315
businv	**0.927**	0.588	0.588	0.588	0.455	0.386	0.370	0.294	0.490	0.275	0.370	0.261	0.588	0.289	0.588
centralia	0.984	0.909	**1.000**	**1.000**	**1.000**	**1.000**	0.909	**1.000**	**1.000**	0.763	0.846	**1.000**	**1.000**	0.556	0.763
children_per_woman	**0.879**	0.678	0.663	0.712	0.405	0.344	0.551	0.525	0.637	0.310	0.504	0.246	0.637	0.500	0.507
co2_canada	0.851	0.544	0.856	**0.924**	0.479	0.642	0.875	0.867	0.670	0.482	0.542	0.569	0.872	0.681	0.361
construction	**0.933**	0.696	0.709	0.709	0.410	0.602	0.709	0.634	0.709	0.324	0.340	0.185	0.709	0.523	0.696
debt_ireland	0.974	0.760	**1.000**	**1.000**	0.892	0.958	0.980	**1.000**	**1.000**	0.469	0.748	0.824	**1.000**	0.538	0.469
gdp_argentina	**0.968**	0.889	0.947	0.947	0.583	0.818	0.889	0.800	0.947	0.615	0.452	0.615	0.947	0.421	0.824
gdp_croatia	**1.000**	**1.000**	0.824	**1.000**	0.583	**1.000**	0.824	0.583	0.824	0.824	0.824	0.400	0.824	0.167	0.824
gdp_iran	**0.921**	0.696	0.652	0.862	0.492	0.620	0.824	0.734	0.808	0.652	0.737	0.636	0.808	0.576	0.652
gdp_japan	**1.000**	**1.000**	0.889	**1.000**	0.615	0.667	**1.000**	0.500	0.889	0.889	0.889	0.222	0.889	0.222	0.889
global_co2	0.625	**0.929**	**0.929**	0.889	0.458	0.667	**0.929**	0.667	**0.929**	0.463	0.547	0.293	**0.929**	0.250	0.846
homeruns	**0.933**	0.812	0.829	0.829	0.650	0.650	0.829	0.829	0.812	0.723	0.397	0.661	0.812	0.664	0.659
iceland_tourism	0.652	0.947	0.947	0.947	0.486	0.391	**1.000**	0.486	0.643	0.220	0.667	0.200	0.947	0.200	0.947
jfk_passengers	**0.978**	0.776	0.776	0.776	0.650	0.602	0.651	0.437	0.776	0.354	T	0.491	0.776	0.437	0.723
lga_passengers	0.885	0.561	0.620	0.704	0.563	0.606	**0.892**	0.526	0.537	0.366	T	0.592	0.537	0.674	0.535
measles	0	**0.947**	**0.947**	**0.947**	0.486	0.118	0.080	0.281	0.153	0.391	F/T	0.030	**0.947**	0.041	**0.947**
nile	**1.000**	**1.000**	**1.000**	**1.000**	0.800	**1.000**	**1.000**	0.824	**1.000**	0.824	0.667	**1.000**	**1.000**	**1.000**	0.824
ozone	0.857	0.776	0.723	0.857	0.778	0.750	**1.000**	0.667	**1.000**	0.723	0.651	0.429	**1.000**	0.286	0.723
quality_control_1	**1.000**	**1.000**	**1.000**	**1.000**	0.667	0.667	**1.000**	0.667	**1.000**	0.500	0.286	0.667	**1.000**	0.667	0.667
quality_control_2	**1.000**	**1.000**	**1.000**	**1.000**	0.667	**1.000**	**1.000**	**1.000**	**1.000**	0.750	.429	**1.000**	**1.000**	**1.000**	0.750
quality_control_3	**1.000**	**1.000**	**1.000**	**1.000**	0.766	0.571	**1.000**	**1.000**	**1.000**	0.667	T	0.800	**1.000**	**1.000**	0.667
quality_control_4	**0.974**	0.810	0.873	0.787	0.561	0.658	0.726	0.658	0.780	0.780	T	0.241	0.780	0.608	0.780
quality_control_5	0	**1.000**	**1.000**	**1.000**	0.500	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.500	**1.000**	**1.000**	**1.000**	**1.000**
rail_lines	0.909	0.846	0.846	**0.966**	0.889	**0.966**	**0.966**	0.800	0.846	0.537	0.730	0.615	0.889	0.205	0.537
ratner_stock	0.933	0.776	0.824	0.868	0.559	0.396	0.776	0.754	0.824	0.280	T	0.203	0.824	0.378	0.571
robocalls	**0.979**	0.800	0.966	0.966	0.750	0.862	0.966	0.966	0.966	0.636	0.846	0.714	0.966	0.714	0.636
scanline_126007	0.887	0.710	0.920	**0.921**	0.829	0.906	0.870	0.838	0.889	0.644	T	0.649	0.889	0.818	0.644
scanline_42049	**0.977**	0.485	0.879	0.962	0.889	0.713	0.910	0.908	0.910	0.269	T	0.460	0.910	0.650	0.276
seatbelts	0.659	0.824	**0.838**	0.683	0.583	0.735	0.683	0.621	0.683	0.452	0.383	0.563	0.735	0.583	0.621
shanghai_license	**0.979**	0.966	0.868	0.868	0.605	0.600	0.868	0.465	0.868	0.532	0.389	0.357	0.868	0.385	0.636
uk_coal_employment	F	F	F	F	0.617	F	0.513	0.513	F	**0.639**	F	F	F	F	0.513
unemployment_nl	0.820	0.742	**0.889**	0.876	0.592	0.747	0.755	0.744	0.788	0.566	F/T	0.628	0.788	0.801	0.566
us_population	0.636	**1.000**	0.889	**1.000**	0.615	0.232	0.471	0.276	0.500	0.159	T	0.889	0.889	0.113	0.889
usd_isk	**0.914**	0.785	0.704	0.785	0.678	0.674	0.785	0.601	0.657	0.489	0.510	0.462	0.678	0.636	0.489
well_log	0.814	0.336	0.914	0.832	0.743	0.822	**0.928**	0.776	0.873	0.149	T	0.923	0.873	0.832	0.237
Average F1-measure (1D)	**0.845**	0.739	0.798	0.822	0.596	0.651	0.784	0.657	0.766	0.482	0.354	0.517	0.797	0.517	0.599
Multidimensional															
apple	**0.949**			0.916	0.445		0.745	0.634			F/T				0.594
bee_waggle_6	0.657			**0.929**	0.481		0.233	0.634			0.245				**0.929**
occupancy	**0.953**			0.919	0.735		0.932	0.812			F/T				0.341
run_log	0.994			**1.000**	0.469		0.990	0.909			0.380				0.446
Average F1-measure (ALL)	**0.871**	n.a.	n.a.	0.855	0.604	n.a.	0.797	0.683	n.a.	n.a.	0.343	n.a.	n.a.	n.a.	0.61
WINS (ALL)	16	0	2	2	0	0	3	0	0	0	0	0	0	0	0
DRAWS (ALL)	6	9	8	11	1	6	8	4	9	1	0	3	8	2	0
LOSES (ALL)	12	25	24	21	33	28	23	30	25	33	34	31	26	32	34

of 0.67 with references of 0.64 (WL), and 0.34 (BS).

6.1.1.3 Segmentation Benchmark

In order to compare the proposed method with other state-of-the-art approaches, we used a benchmark provided by the Alan Turing Institute [26] (Dataset 9, ATCPD, see Section A.1.9). The performance was evaluated by change point event detection in each time series available, summarized in Figure 6.1, where each referenced method applied its best score in the benchmark (see Tables 6.3 and B.7).

As unfolded in Figure 6.1, the critical distance diagram ranks the proposed method second, suggesting no significant difference in performance among methods that only work in uni-dimensional datasets. The global average F1 measure of the proposed method is 0.87 for both uni and multidimensional datasets, with a higher F1-score than all other

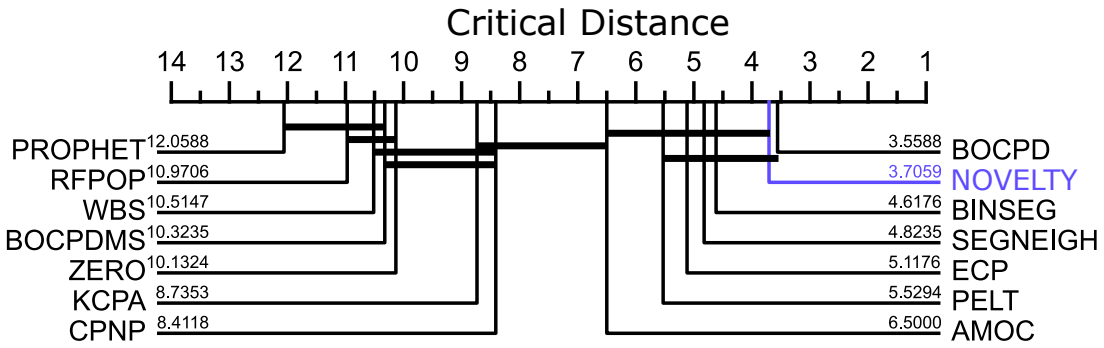


Figure 6.1: Critical distance diagram comparing the methods used in [26] (except *RBOCPDMS*) and the *novelty function* on Dataset 9 (see Section A.1.9). The performance measure corresponds to the F1-score for all single-dimension datasets of the benchmark, except for the ones identified in Table 6.3 with a gray background. A thick horizontal line groups a set of classifiers that are not significantly different in the statistical test [84].

methods in 16 of the 34 datasets. In contrast, the *BOCPD* method had an overall of 2 wins against all other methods. The two null scores are because no change point was supposed to be found in the corresponding time series. The results obtained in this benchmark restate that our proposed method is promising, having a performance that competes with several state-of-the-art methods in the problem of novelty segmentation. It should be stressed that the proposed method applies to multidimensional time series, while two of the best-ranked methods in Figure 6.1 do not. In addition, the proposed method retrieves not solely segmentation points but also higher-level information as periodic changes and cross-segment similarity measures, which is an advantage over the *BOCPD*.

6.1.1.4 Time Complexity

In terms of computation time, the algorithm performs (1) a sliding window to extract features, which is $O(n)$ complexity and (2) performs the dot product between matrices, which is traditionally $O(m^2 n)$ (recall that r is the number of features and n is the size of the inputted signal). Finally, the correlation of a kernel on the SSM's diagonal has a complexity of $O(nM^2)$ (recall that $M = 2L + 1$, being the size of the sliding kernel).

The sliding window to extract features has an $O(n)$ time complexity. The dot product between matrices has a conventional $O(m^2 n)$ time complexity. Expressly, m and n represent the number of features and the inputted signal's size, respectively, in our proposed method. The correlation computation of a kernel on the SSM's diagonal has a complexity of $O(nM^2)$, where the sliding kernel's size M equals $2L + 1$.

6.1.1.5 Distance of True Positives to Ground Truth

From the results presented in the previous Section, we showed that the proposed method is capable of identifying changes in signals. However, how precise is the method in detecting these events? For this, we measured the distance from the TP events and the

corresponding ground truth. A more detailed explanation of these results are showed in Appendix B (Section B.2.2).

The distance values, in seconds, show that, typically, the proposed method, with the parameters used, is less precise than the other methods in most datasets. It is a clear limitation at this stage, but we identified several reasons for this:

- ***search timescale***: The larger the window size, the less precise is the detection, which makes sense, considering that the resulting novelty function will be more smoothed and peaks will be detected with less precision;
- ***signal size***: The larger the signal, the more difficult it becomes to compute the SSM. This results in either reducing the overlap size, or increasing the window size. In both cases, the resulting novelty function will be more smoothed;

The mentioned reasons influence the parameter selection that affects the precision in the detection. This can be corrected if using more appropriate overlap size (higher overlap) and detection strategy (multi-timescale search and different peak strategy).

6.1.1.6 Discussion of Novelty Segmentation Results

Several parameters affect the detection results of desired patterns, especially the window size, the overlap percentage, and the kernel size, which influence visual outputs and the novelty function. These parameters can be explained with the analogy of a camera:

- The window size works like the *zoom function*, defining the scale of interest in the time series. Larger windows, corresponding to lower *zoom values*, allow the similarity calculation of longer *subsequences*, while smaller windows, serving like a *zoom-in* function, search for local details and unobtrusive changes.
- The overlap percentage, working as a down-sampler of the time series, is the *camera sensor*, which determines the image's pixel resolution. A full resolution of the SSM is only achieved with total overlap, and the lower the overlap percentage, the less accurate are the highlighted changes.
- The sliding kernel's size concerns the novelty function's sharpness of the detected changes. The larger it is, the smoother the output function will be. Potentially, the kernel size can be scaled to the window size, even with a slight accuracy decrease.

With enough computational resources available, the overlap percentage can be maximized so that the SSM can mirror the full details. Admittedly, such an operation is not necessary for many real applications, but reduces the variables to facilitate other parameters' tuning experiments, which is one of our subsequent research topics. The computational resource, i.e., the memory bandwidth and the calculation time (see Section 6.1.1.4) is a limitation in this current stage, since the SSM increases exponentially with the

increase in the time series size. We ascertained that downsampling the time series with a lower overlap percentage is a valid option, advanced by a hierarchical search strategy for addressing the memory limitation, as exemplified in the walking-series instance in Section 4.4.1. Another potential efficiency-enhancing solution is to only compute the SSM's central diagonal with the kernel size corresponding to the interest areas of segmentation borders, which obtains efficiency gains in exchange for the sacrifice of periodicity and similarity measures between *subsequences*.

A reasonable intuition based on the understanding of signal characteristics should help configure the parameters mentioned above that are fundamental for computing the SSM and the novelty function, as Figure 6.2 demonstrates a starter example for segmentation purposes. The upper part of Figure 6.2 draws different SSMs on the same ECG record (A) from Dataset 2 (ECG1, see Section **??**) computed with sequentially larger window lengths from 0.01 to 2 seconds. The appropriate window length depends on the purpose of the search:

- If a small window length, e.g., 0.05 seconds, is chosen, the novelty function will mostly detail changes within a heartbeat.
- If opting for a window length approximately equal to the ECG's *PQRS* complex, each transition between complexes will be projected.
- If even larger windows are applied, e.g., 1 or 2 seconds, the jump artefact will be more significant on the SSM and be spotlighted on the novelty function. Hence, in such a case, the window length of 1 second should be appropriate for segmenting clean versus noisy ECG signals.

When using the same window length on all the other records of the same dataset, the SSM is expected to highlight the same regions of interest. Figure 6.2(bottom) characterizes that parameters can be identical when working on the same data type and purpose, but the peak selection on the novelty function is not a matter of convention, which depends on the preset threshold.

The threshold used to determine which peaks are considered as points of interest is not relevant to the SSM calculations but closely related to event detection and automatic segmentation. If the data is a black box, the choice of threshold is a matter of observation and speculation. With informed knowledge of the data, the threshold can be predetermined and experimented with rules based on ranking the detected peaks from highest to lowest:

- Set the total number (or quantity range) of points of interest as the threshold. E.g., an ECG series with x heartbeats; an ACC series with y recorded gaits of walking activity.
- Count the total number of peaks and specify a percentage as the threshold. This work takes such an approach because of the diverse datasets and signals involved.

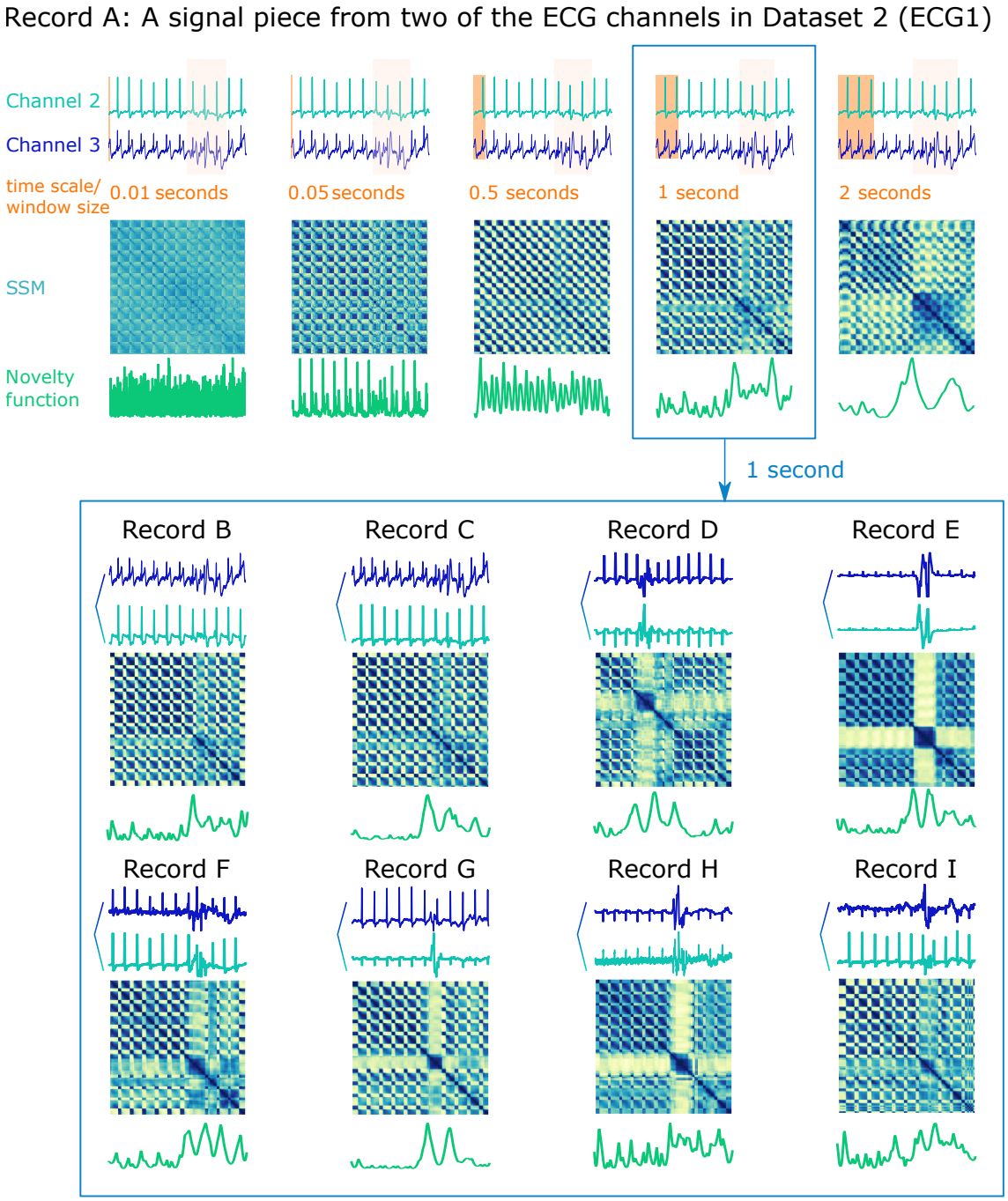


Figure 6.2: Illustrative example of window length intuition on records of Dataset 6 (ECG1, see Section A.1.6). Top: different SSMs on the same ECG record A computed with sequentially larger window lengths from 0.01 to 2 seconds. The novelty functions are calculated with a kernel size equal to the window size and an overlap of 95%. Bottom: The 1-second window length is further applied as an example to indicate that parameters turned in the representation experiments can be generalized to all other records of the same dataset (B-I) to compute their corresponding SSM representations and novelty functions.

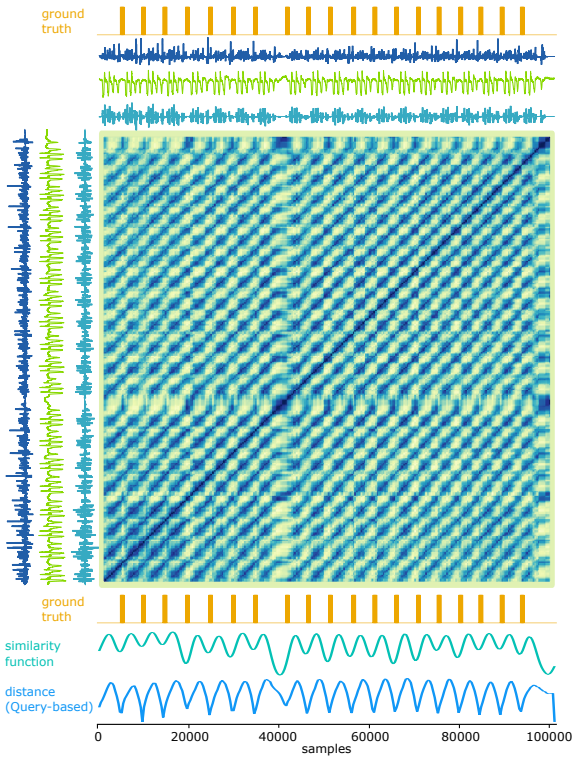


Figure 6.3: Computing the SSM of the periodic signal from subject 1, activity 0 (see Section A.1.11). The ground truth is shown in orange (the first signal). The similarity function and distance profile from the query-based search are showed below the SSM. The SSM was computed with a window size of 5290 samples, and an overlap of 85%.

- Add sliding windows to the novelty function based on the known periodicity information of the time series, and define the number of points of interest expected to be present in each window as the threshold.

6.1.2 Periodic and Query-based Segmentation

The detection of periodic onset was experimented on Dataset 11 (*CSL*, see Section A.1.11). It is a database with exclusively periodic data of activity signals. Twenty participant performed seventeen different activities, which are recorded in a multimodal time series. We exemplify the usage of a periodic search with the similarity function and the query-based segmentation on the SSM. These experiments were evaluated based on ground-truth annotations of a push-button from the dataset, with the metrics previously explained (see Section 2.8.2). The ground truth indicates the onset of the activity. Results are presented in Table 6.4.

Using the similarity function enhances the periodic nature of the time series when computed (see Section 4.3.2). The results are displayed for all activities, summing *TP*, *FP* and *FN* of all participants. The periodic search was made using the window size of one period. The tolerance window was half the period length. The *similarity function* was smoothed by a smoothing function (see Section 2.2.2) with a window size equal to the

Table 6.4: Results in segmenting periodic data from dataset 11 (see Section A.1.11). The results were computed with two different approaches on the SSM: *similarity function* and query-based search. *TP*, *FP*, *FN* and resulting *F1-score* were the metrics used. The total number of these metrics is presented on the last row.

	Similarity function				Query-based search			
Activity	TP	FP	FN	F1-score	TP	FP	FN	F1-score
0	348	37	33	0.91	375	15.0	6	0.97
1	349	36	30	0.91	375	28.0	4	0.96
2	323	39	58	0.87	381	48.0	0	0.94
3	358	27	23	0.93	381	37.0	0	0.95
4	335	17	39	0.92	372	35.0	2	0.95
5	323	75	59	0.83	381	51	1	0.94
6	365	4	16	0.97	380	13	1	0.98
7	354	28	27	0.93	380	15	1	0.98
8	336	25	44	0.91	364	35	16	0.93
9	333	42	27	0.91	354	10	6	0.98
10	341	31	40	0.91	379	17	2	0.98
11	324	15	36	0.93	351	20	9	0.96
12	682	74	52	0.92	718	68	16	0.94
13	770	30	766	0.66	685	122	851	0.58
14	322	50	40	0.88	353	41	9	0.93
15	344	38	17	0.93	360	29	1	0.96
16	584	60	126	0.86	659	26	51	0.94
Total	6791	628	1433	0.89	7248	610	976	0.93

window size used to extract the features. The smoothing function helps in making the valley detection on the resulting similarity function easier and preventing unwanted *FP*.

The results presented in Table 6.4 show that the similarity function can be used to segment periodic signals with a low rate of *FP* and *FN*. The overall F1-score is 0.89, being a good result but with margin for progress. The similarity function highlights this periodic behaviour, but the detection of periods is dependent on the ability to detect the relevant valleys. This process requires either a thresholding method (such as the one used on the novelty function) or the detection of all present valleys. Of course, using a valley/peak detection method with additional parameters, such as the distance between valleys/peaks, are plausible considering the nature of the problem. In this case, we used the simple approach of smoothing the signal and detecting all present valleys, with the risk of increasing *FP* and/or *FN* counts depending on the smoothing factor. Therefore, a better valley detection procedure can improve the current results. *Activity 13* is an exception with a specially high rate of *FN*, which is due to the nature of the ground-truth annotation. Figure 6.3 illustrates an example of the periodic detection with the similarity function and the query-based search approach.

Although this method was mostly successful in segmenting the signal into distinctive periods. It does not always provide the moment where the *path* starts on the SSM. This

moment should be the one detected to capture the true starting point of a period. The fact that we are not able to find the exact moment the *path* starts means that an offset is present on the cycle detection. This is visible in Figure 6.3, where the valleys of the similarity function do not perfectly match the ground truth, when the *paths* of the SSM do. In the future, a different approach should be considered to segment the *paths*. Nevertheless, the SSM can still be used to segment periodic signals with the help of the analyst, by interactively selecting a segment of the signal as a query, or simply by indicating the index that corresponds to the beginning of the period, which is computed by query-based search with the SSM' columns/rows, as explained in Section 2.6.3. This query-based search was also used in the same dataset and the results can be found in Table 6.4 and Figure 6.3. The query selected as template was the first period of the signal and the tolerance window equal to half the period size. The results show an expected improvement in the results, from 0.89 to 0.93 for the *F1-score*. In this case, it is notorious that the *FN* rate is much smaller than the previous case, which makes sense because a valley would always be present where the query overlaps each period.

It is also important to mention that this problem is not particularly complex for this dataset, considering that it is a controlled dataset with one periodic signal. But our intention with this experiment was to demonstrate the ability of the SSM in (1) highlighting if a segment of the multimodal signal is periodic with the presence of *paths*, and (2) segmenting the periods of this segment without any prior knowledge on the signal besides the window size, which is more flexible. Solely finding periodicity can be done with other methods, but in this case the SSM shows all variations on a time series, with novelty and periodicity.

6.1.3 Information Retrieval in Occupational Data

The current information era, in conjunction with technological developments, is promoting a shift in the way industries and companies manage and perform their work. With the inclusion of sensing devices, dashboards and machine learning algorithms, many tasks can be better evaluated with direct quantitative measures. This can be made for machines to prevent breakdowns and optimize their performance, but can also be used for humans, specially to evaluate their occupational health, and prevent work-related disorders. This prevention implies a specialized intervention to perform changes in the workplace, the collaborator's training process and/or other organizational strategies. The process of deciding a need for intervention implies a previous risk evaluation of the collaborator's or team's routines. However, before the actual ergonomic risk evaluation, a considerable number of steps in preparing the data are needed.

As we mentioned in Section 1.3, data preparation is an important part of any analysis or task that requires further deployment in supervised methods, namely the time-consuming segmentation and labelling processes. Hence, tools capable of segmenting time series to support and ease the labelling process of motion and posture data are highly valued and highly necessary for the risk analysis and the deployment of semi-supervised and/or

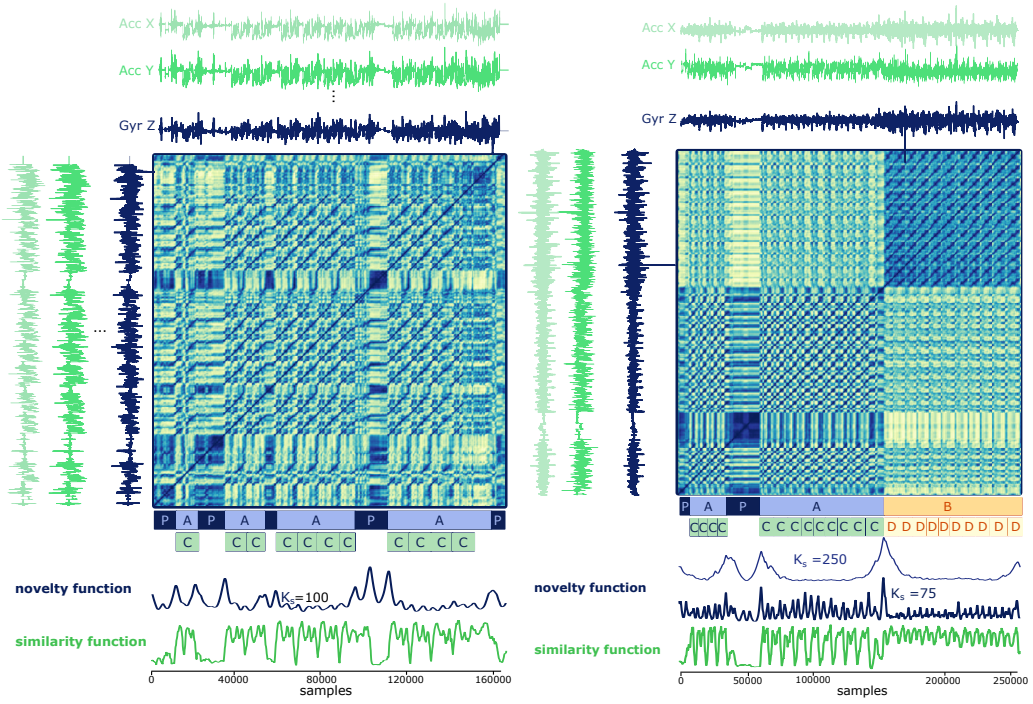
supervised methods. In this case, motion data in real scenarios are even more complex since it is highly rich and diverse in behaviours. For instance, although cyclic and repetitive activities performed in industrial scenarios are mostly consistent from cycle to cycle, perfect conditions cannot be met all the time. Eventually, the working process can inadvertently be stopped or delayed. In addition, ergonomic risk assessment methods have to analyse each working cycle, which means that a previous segmentation of each one of these instances has to be made. Sub-activities that make the sequence of actions comprehended in the working cycle can also be sub-segmented for further recognition and association with risk measures. For instance, the *ergonomist* might have an interest in understanding which actions from the working cycle have a significant association with a high or low risk measure, and adapt the working station with intervention strategies that could help prevent future disorders.

In this Section, we explore a few examples in using the SSM for information retrieval in occupational motion data of collaborators performing specific workstations. The data was recorded at the *Autoeuropa Volkswagen* production line (Dataset 13, *VAOD*, see Section A.1.13). The information provided by the SSM supports the analyst in identifying the following events:

- **Active working periods:** *Active and non-active segments* indicate sub-sequences of the time series where the collaborator was performing the cyclic working tasks or not (e.g.: stop on the working line that leads to a pause in the working activity). With this information, we can focus the attention to active working periods, which are the segments of interest to perform risk analysis. For this, the *novelty function* can provide the moments of change between active and non-active regimes, as we will show further;

- **Working cycles:** Sub-sequences of highly similar sequences of movements that are being repeated in time. This type of segmentation pre-assumes that the time series under analysis will be mostly (or entirely) defined by the repetition of a motion. The illustrated examples show only tasks that are cyclic. The *similarity function* highlights moments with periodic behaviour.

- **Sub-cyclic segments:** Sub-cyclic segments inside each working cycle can be compared to evaluate how much it contributes to the risk score of a workstation. Segmenting the same portion of the working cycle over time can provide relevant comparative measures for decision making or intervention in the production line (e.g. it can be decided that the portion of the cycle has to be change to reduce the exposure to risk factors). The usage of a query-based mechanism with an example of the *subsequence* is used to demonstrate the findings.



(a) (b)

Figure 6.4: Illustrative examples that show how to use the SSM for information retrieval in motion data of occupational scenarios. In both figures, a set of signals, presented as *Acc X, Y* and *Gyr Z* show a sample of the data used (Acc is the ACC and Gyr is the GYR data). The data from the wrist and elbow were used for this analysis. The segments are illustrated and labelled below the SSM with ***P*** - *interruption/pause in the production line*, ***A*** - *active working period*, ***C*** - *working cycle*, ***B*** - *active working period for workstation 2* and ***D*** - *working cycle 2*. The novelty and similarity functions are presented to show the match with the segmentation labels. In (a), the signal represent a collaborator performing one workstation with interruptions, while in (b) a collaborator working in two different workstations.

6.1.3.1 Novelty and Periodic Segmentation in Occupational Data

In Figures 6.4a and 6.4b we present two different examples of data acquired while the collaborators were working. In Figure 6.4a, collaborator 1 was performing the cyclic task 1, while in Figure 6.4b, collaborator 2 was performing two different tasks (1 and 2). It is clear in both Figures that the SSM provides a visual feedback that highlights moments where the signal change (*blocks*), as well as where cycles occur (*paths*). In the first case, the collaborator waited (**P**) to start the task (**A**), up until the collaborator was interrupted several times by a stoppage on the production line to then continue his/her tasks. The novelty function is able to detect all of these main transitions, while the similarity function highlights where the working cycles (**C**) are.

In the second example, collaborator 2 was performing the same workstation (**A**) as collaborator 1 but shifted to another workstation (**B**). The SSM indicates well when transitions occur, namely where the collaborator was interrupted while performing workstation

A, and when shifting to workstation **B**. This is highlighted with the novelty function. The cycles are also visible with *paths*, for both workstations. The similarity function highlights well where the working cycles (**C** and **D**) happen. However, the search for working cycles should be performed after a first segmentation between homogeneous segments in order to have better results.

It is also relevant to mention that when the collaborator is in non-active periods (**P**), the *similarity function* has a very low value. As we mentioned in Section 4.3.2, a low similarity function value also corresponds to being very dissimilar to all the other *subsequences* of the signal. In this case, it makes sense, considering that the majority of the signal is related with cyclic activities, being **P** instances less frequent and therefore less similar with all the other *subsequences*. This happens in both examples.

The *novelty* and *similarity* functions were used with the purpose of detecting the listed events (see List 6.1.3) on Dataset 13 (*VAOD*, see Section A.1.13). For the sake of brevity, the results for both item one and two of the list are displayed in Appendix B (see Tables B.8 and B.9, respectively). The first step involved detecting changes between active and non-active periods of the working activity, which was performed with the novelty function. The periods were described as active or non-active based on a threshold on the similarity function. The results show the ability to segment the signals with an overall F1-score of 0.96 (see Table B.8). After identifying the active periods, working cycles were segmented. From the 157 working cycles, 154 cycles were detected (see Table B.9).

6.1.3.2 Query-based Segmentation in Occupational Data

Another aspect of occupational risk evaluation in industrial scenarios is to compare the exposure to risk factors of sub-segments of the working cycle. This strategy could support professionals in identifying specific sequences of sub-activities that occur in-cycle, and search them over the entire set of working cycles for comparative purposes. In Figure 6.5 we show an example where the subsequence search matches exactly two distinctive in-cycle patterns (1 - blue; and 2 - green) of the working activity. Using these subsequences as a query template, we can match their onset over the successive cycles by finding the major valleys of the resulting distance functions (see Equation 4.8).

6.2 Pattern Search with SSTS

In order to show the value of using a symbolic representation of the time series and search with regex on that representation, we provide six examples of how to use SSTS for pattern search on time series and compare the text complexity between the python implementation and the query used, with Halstead measures.

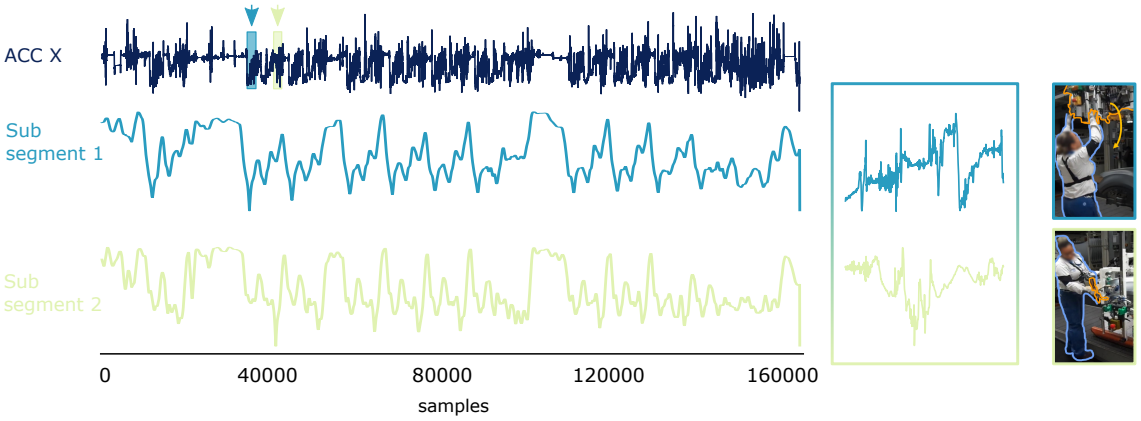


Figure 6.5: Illustrative example that shows how to use the SSM for information retrieval in motion data of occupational scenarios. The query-based search with a template is made on the SSM of the signal. This search procedure highlights where the query re-occurs on the signal with the distance functions (*subsegment 1* and *subsegment 2*). The templates are highlighted on the signal (Acc X) and on the right of the resulting subsegments. The small thumbnails next to the template illustrates the actions being performed on the cycle by the collaborator (highlighted in blue) and the tool being used (highlighted in orange). On the first subsegment, the collaborator was reaching over the head and pulling down the working tool, while on the second subsegment, the collaborator was screwing the pieces on the side of the car.

6.2.1 Illustrative Evaluation of SSTS in Various Application Scenarios

We will present six examples (three more detailed): the second and third examples consist of simple problems that require the detection of local maxima and minima. The sixth example is a more challenging task, which requires the usage of more sophisticated mechanisms of the regex module. The examples are listed as follows(*Example 1* is omitted since it was discussed in the previous section):

- Example 2 - Step Detection: Detect the instants in time when a right or left heel floor contact is achieved during a normal straight walk from the magnitude of an accelerometer signal. This information can be used to create a step detector based on accelerometer data. Since the sensor is located on the right pocket, global minima will indicate the right heel contact and local minima will correspond to left heel contact;

- Example 3 - Segmentation of the systole of the ABP wave: The dicrotic notch corresponds to the physiological event of the aortic valve closure, which triggers the increase of the aortic pressure and signifies the end of the systolic phase. In this problem, it is asked to segment the systolic phase of the ABP wave;

- Example 4 - Electrocardiogram peak detector: Detect all the major peaks from an ECG record;

- Example 5 - Straight Line Trajectory Tracking: The automatic detection of trajectory features, such as straight walks and turns can be used to evaluate trajectory anomalies from a vector of cartesian coordinates in 2D-space;
- Example 6 - Since the first 5 s of the lifting exercise are unstable, the problem involves detecting the first stable lifting step, which occurs approximately 5 s after the beginning of the exercise from an accelerometer signal.

For the sake of brevity, only three of the examples are presented.

6.2.1.1 Example 2 - Step Detection in accelerometer signals

In this particular case, only the detection of the right heel contact will be discussed, whereas the left heel contact example's solution is presented in Table 6.5.

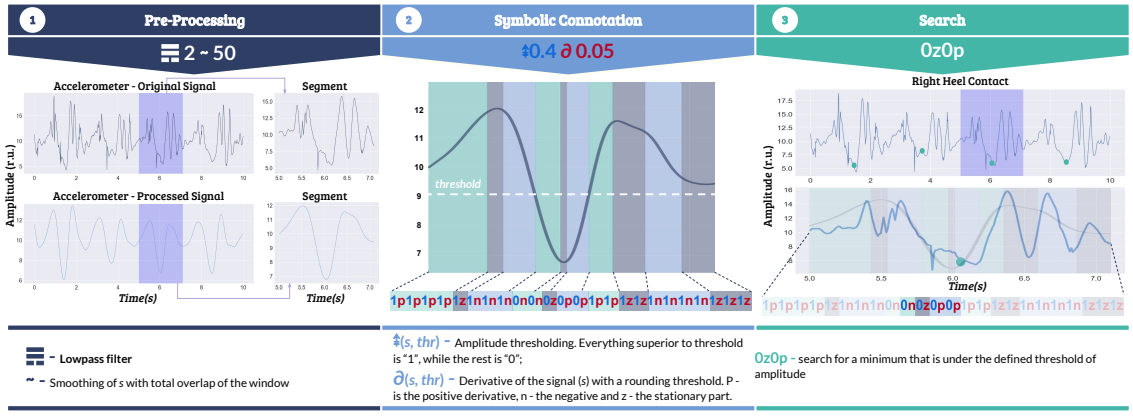


Figure 6.6: Example 2 - Solution pipeline of Step Detection example. At the bottom of the figure are summarized the operators and methods used in each step. In the symbolic connotation step, the alternation between the methods is made with colors (Blue - A - Amplitude Comparison, Red - D1 - First Derivative). A dotted line is shown to represent the threshold level. Each color band in the signal corresponds to a specific primitive. A positive match is highlighted with green in the search step.

The signal is initially pre-processed to ease the identification of the subject's steps. This is achieved using a low pass filter (**LP**). The result is presented in the second image of the first step in Fig. 6.6. A highlight is also present and delineates a segment of the signal that has a minimum peak, which corresponds to right heel contact. The string representation of this segment is depicted in the second step.

The string is a sequence of primitives composed of two connotation methods (**A** in blue, **D1** in red). Based on these methods, the samples of the signal with amplitudes greater than *40*% of the range amplitude are transcribed into 1, while the remaining turn into 0; the slopes are converted into p, z and n, when rising, being stationary and falling, respectively.

The solution involves detecting each minimum with an amplitude inferior to the threshold level. The morphological representation of a minimum can be reduced to a

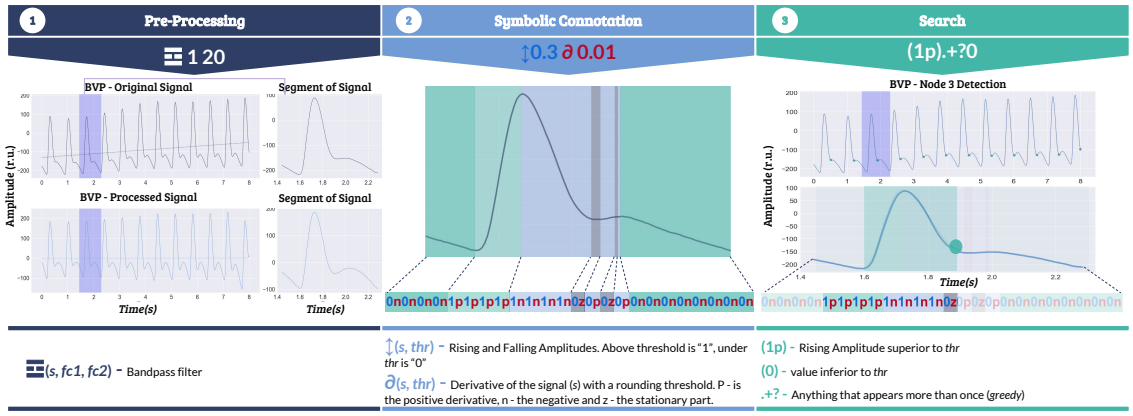


Figure 6.7: Example 3 - Solution pipeline of Dicrotic notch detection example. The operators or methods used in each step are summarized at the bottom of the figure. In the symbolic connotation step, the alternation between methods is represented with colors (Blue - AD, Red - D1). For each color band in the signal, there is a corresponding primitive. The match is highlighted with green in the search step.

negative slope followed by a positive one, which in the symbolic representation is defined by the second connotation method as the transition from `n` to `p` or `z` to `p`. Regarding the amplitude requirement, it is assigned by the first connotation method, in which any value lower than the threshold level is `0`. The regex used to find this minimum was `0z0p`, which implies that: (1) the amplitude has to be `0`; and (2) the derivative is `z` and then `p`. The detection is highlighted in green in the last plot of Fig. 6.6.

6.2.1.2 Example 3 - Dicrotic notch detection in ABP signals

The ABP waves are morphologically represented by a high positive slope that corresponds to the systolic uptake and ends at the peak of the systolic pressure. After this behaviour, follows the systolic decline, which ends with the aortic valve closure, named the *dicrotic notch* in the signal representation [138].

These types of signals are commonly affected by low-frequency noise that modulates the signal and is contaminated with high-frequency noise of low amplitude. The typical procedure is to bandpass the ABP signal to remove both types of noise. Fig. 6.7 depicts how a band pass filter that cuts frequencies under 1 Hz and above 20 Hz (BP) is applied to the signal to remove both types of noise. The modulation removal is shown with the linear regression in both original and pre-processed signals, which in the first case has a positive slope and after the pre-processing is approximately zero.

In the symbolic connotation step, the reasoning follows the signal morphological description and uses two methods for the symbolic representation. The first, AD (represented in blue), detects all rising and falling slopes that are higher than a specific threshold (in this case 30% of the amplitude range of the signal) and transcribes the sample values to `1`, while the remaining values are converted to 0. The second method (represented in red) uses the derivative, as mentioned in the previous examples.

The first connotation method is necessary to distinguish between the rising slope that occurs at the beginning of the pressure wave and the one after the *dicrotic notch*. With this distinction, it is possible to find the beginning of the ABP wave as a high positive slope (`1p`) and find the *dicrotic notch* when the lower slope starts (`0.`). In order to find this area of the ABP wave, the regex has to start with the first `1p` primitive and end with the first `0.` primitive.

The example is solved with the following regex: `(1p).*?(0.)`. This string means that the search will match anything (represented by "`.*?`") between the first "`1p`" primitive and the first "`0.`" primitive.

6.2.1.3 Example 6 - Stable lifting detection in accelerometer signals

In the previous example, the solution was achieved by searching for one simple transition in the string generated by the sequence of connotation methods. This tool may also be used to solve more complex examples in the same manner.

The next problem involves the segmentation of a lifting step that has occurred 5 seconds after the start of a weight lifting exercise. The example can be solved in two steps: (1) find the start of the exercise, and (2) search for the segment 5s after the start. This rationale can be expressed by combining two distinct symbolic representations of the same original signal, which requires the use of two pre-processing and symbolic connotation sets, in which one is used for the detection of the start and the other to find the desired segment.

Fig. 6.8 demonstrates how the example is solved. In both pre-processing and symbolic connotation steps, a vertical bar `|` separates the methods that are applied for each representation of the same signal. The pre-processing phase uses a sequence of whitening, modulus, and smoothing of the signal for "*Process A*", and a sequence of whitening and smoothing for "*Process B*". The first processing sequence turns the signal similar to a plateau, in which the beginning of the activity is easily identified; whereas, in the second sequence, the signal is smoothed such that each lifting step is well defined.

Regarding the symbolic connotation step, the first method AD (represented in blue) turns all sample values of the first signal that are higher than the threshold level into `1`, while the remaining samples are converted to `0`. The second set is applied to the second signal and uses the derivative of the signal D1(represented in red). Both symbolic connotations merge into a sequence of primitives, in which the first element inspects if the exercise has already started (`1.`) or not (`0.`), and the second infers the sectioning of the lifting steps, that is if the sample of the signal is increasing (`.p`), stationary (`.z`) or decreasing (`.n`).

The regex written to solve the example is `(?<=1.{15000,})(n1p)(.*?)(n1p)`. Decomposing this expression in the two steps of the example results in: (1) `(?<=1.15000,)` and (2)`(n1p)(.*?)(n1p)`. In (1), a *lookbehind* operator (recall rules from Table 2.1 in Section 2.7.4) is used, therefore, the compiler will search ahead of the first match inside the operator, i.e., the search will match the expression "`1.{15000,}`" and search ahead

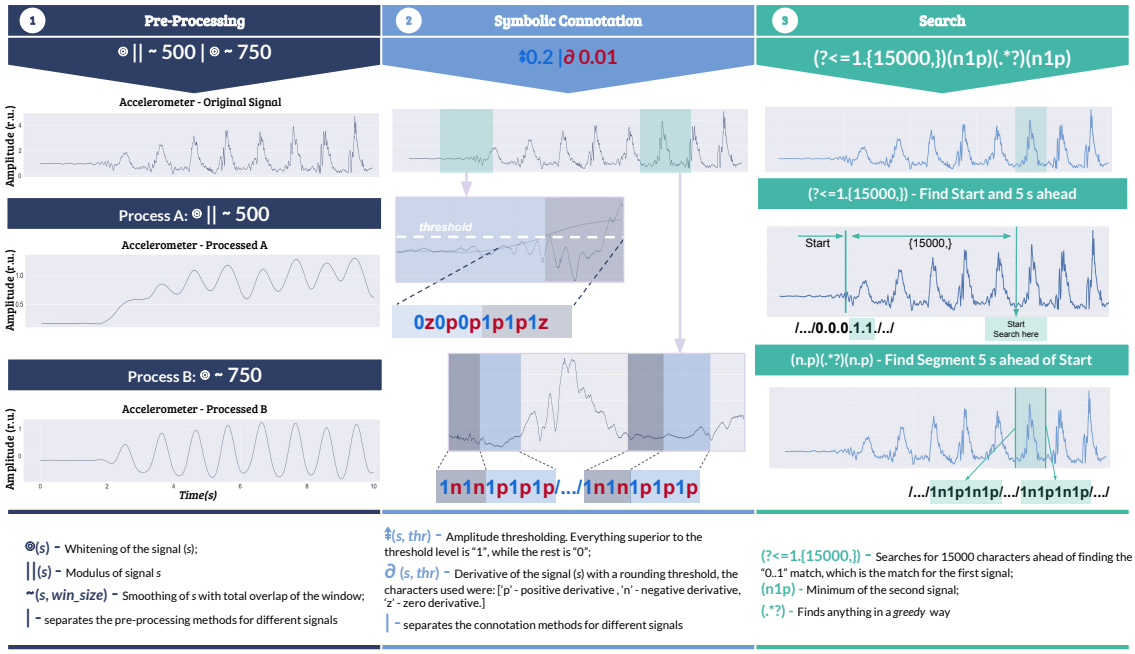


Figure 6.8: Example 6 - Solution pipeline of Stable Lifting Detection. The operators or methods used in each step are summarized at the bottom of the figure. In the symbolic connotation step, the alternation between the methods is made with colors (Blue - **AD**, Red - **D1**). The threshold level is identified with a white dotted line. Each color band in the signal corresponds to a specific primitive. The match is highlighted with green in the "Search" step.

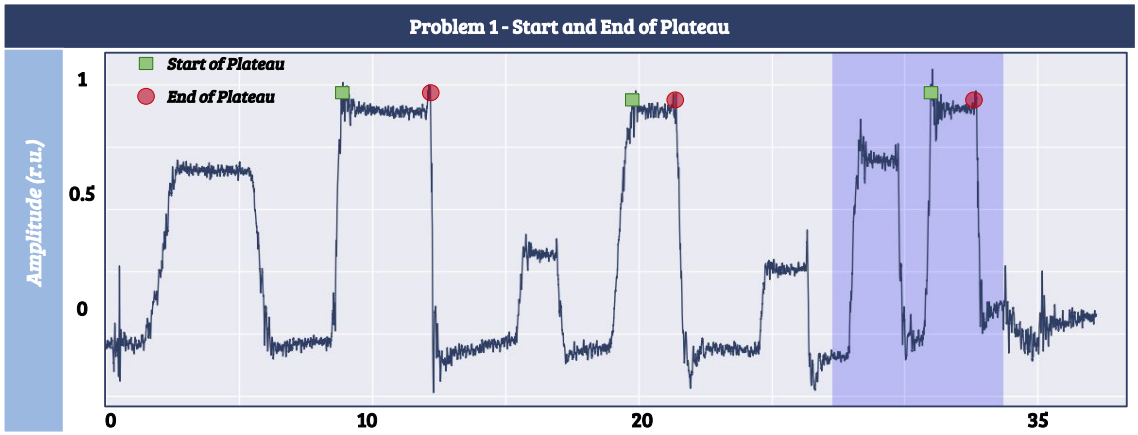
Table 6.5: Summary of the SSTS queries used to solve the six illustrative problems.

Example	Fig.	Pre-Processing	Connotation	Search
Start of Plateau	6.9.a	S 500	A 0.8 D1 0.05	`p1n`
End of Plateau	6.9.a	S 500	A 0.8 D1 0.05	`z1n`
Step Detection (Left)	6.9.b	LP 2 Sm 50	A 0.4 D1 0.05	`0z0p`
Step Detection (Right)	6.9.b	LP 2 Sm 50	A 0.3 D1 0.05	`1z1p`
Dicrotic Notch	6.9.c	BP 1 20	A 0.3 D1 0.01	`(1p).+?0`
Electrocardiogram Peak Detector	6.9.d	BP 5 50	D1 0.01	`pn`
Straight Line Tracking	6.9.e	none	D2 0.05	`z*?`
Stable Lifting Detection	6.9.f	Mag Abs Sm 1000 | Mag Sm 750	A 0.2 | A -0.2 D1 0.01	`(?<=1.15000,)(n.p)(.*?)(n.p)`

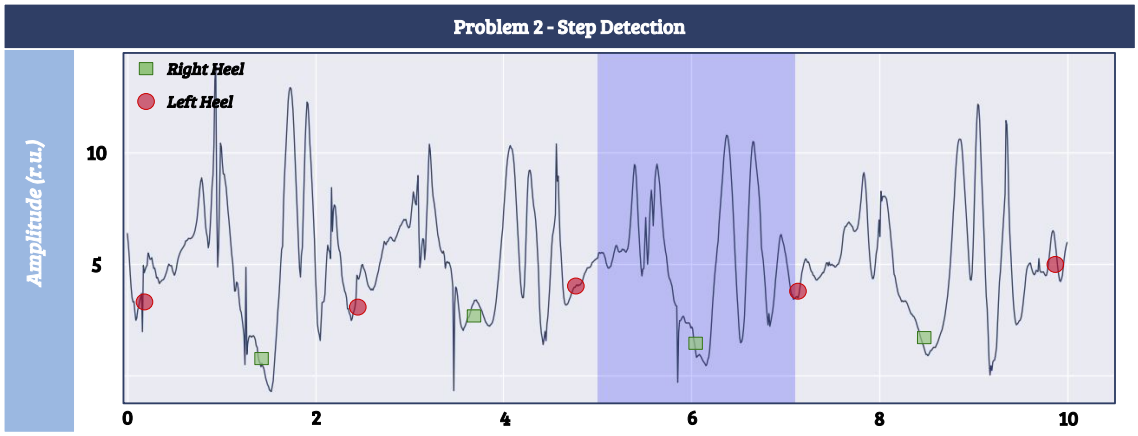
of it. This method aims to handle the temporal dependence between events in the string, in which the number `15000` is calculated by the number of characters that correspond to 5 seconds in the string. This calculation was achieved by multiplying the sampling frequency, the number of connotations in each primitive, and the desired time, in this case: 1000 Hz, 3 connotations, and 5s, respectively.

6.2.1.4 Summary of all examples

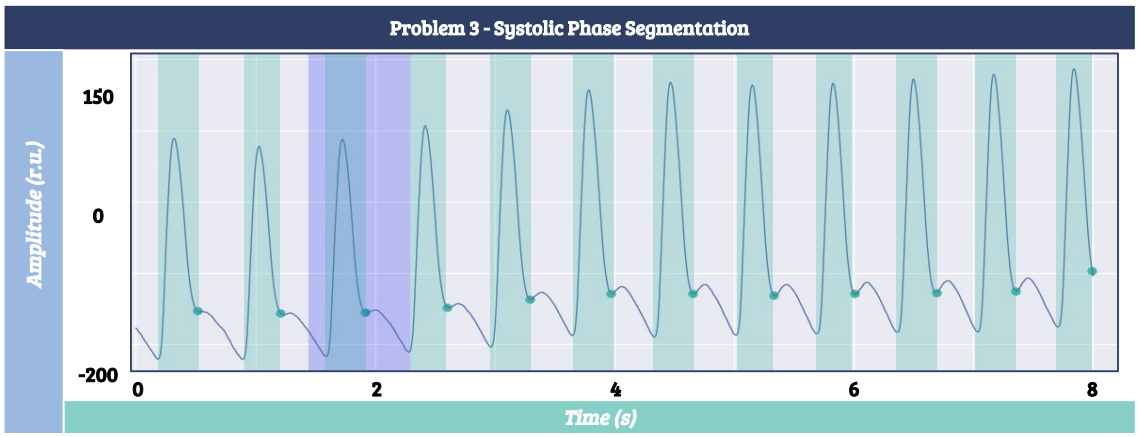
At this point, introductory examples have been explained and can be used to understand the workability of the tool. Table 6.5 and Fig. 6.9 summarize the example's resolution for each step of the pipeline. Additional examples are also included and can be found in the



(a)

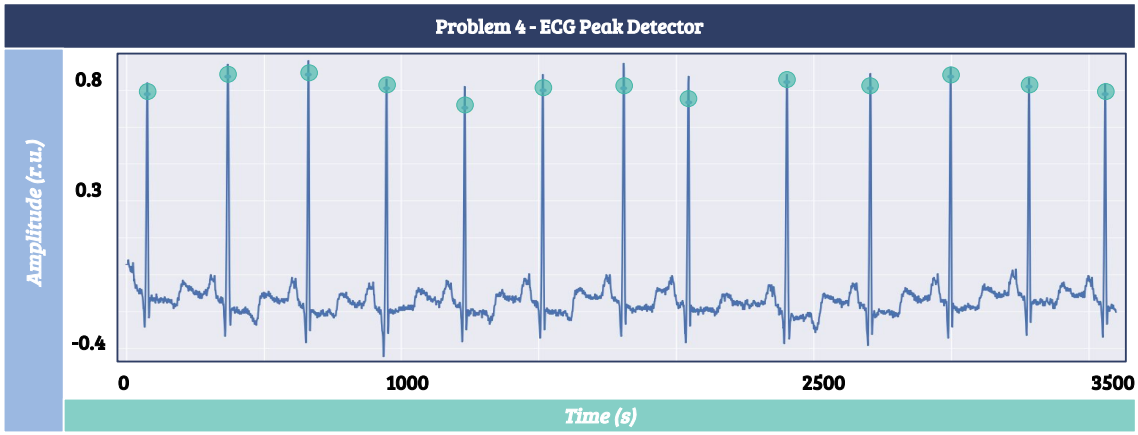


(b)

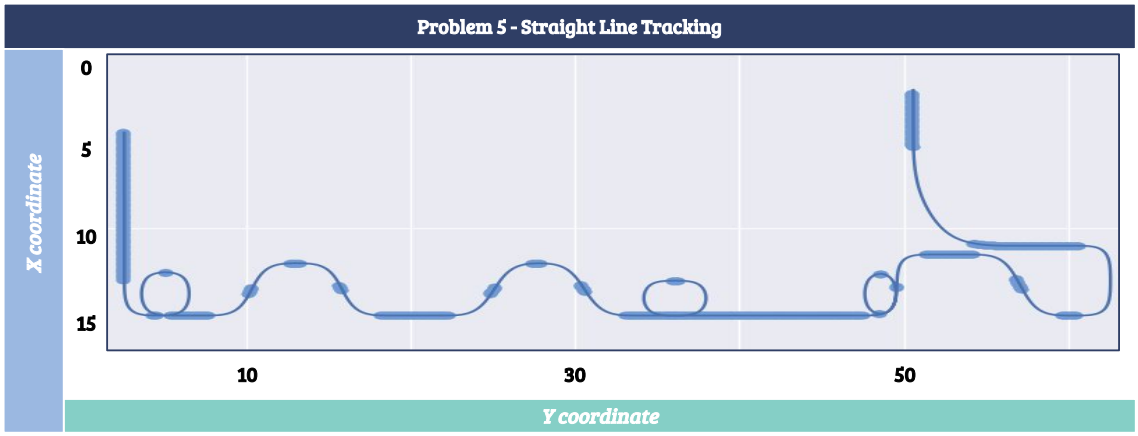


(c)

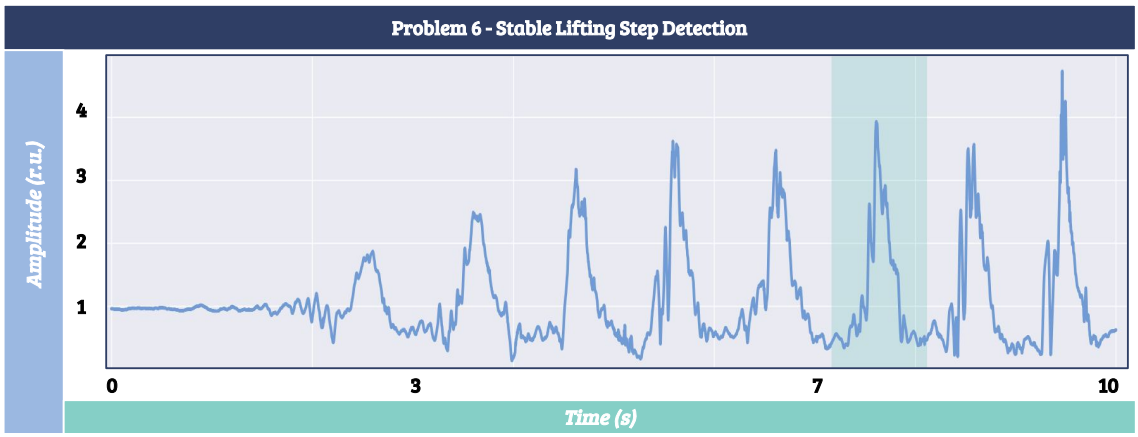
Figure 6.9: Summary of the resolution of the problems 1 (), 2 (Step Detection), and 3 (Segmentation of the systole on the ABP wave), listed in the previous Section with the SSTS.



(d)



(e)



(f)

Figure 6.9: (Continuation of the previous figure). Summary of the resolution of the problems 4 (Electrocardiogram peak detector), 5 (Straight line trajectory tracking), and 6 (Lifting exercise), listed in the previous Section with the SSTS.

GitHub repository of the tool[3]. You can also try out in a web version, available at [4].

6.2.2 Measure of Expressiveness

In order to evaluate parameters of expressiveness with SSTS, we evaluated the legibility and difficulty in generating the solution for the corresponding task. In the programming field, *Halstead* measures are typically used for this purpose. These measures describe the complexity of the script directly from the source code, based in a set of metrics calculated with the number of distinct oprt and oprd, and the Toprt and Toprd (recall Section 2.8.3).

The complexity evaluation will be performed to compare the script generated by the *classical* approach with the *syntactic* approach. Concerning the *classical* approach, the list of operators and operands are selected from the *operator* module from python [50], which includes the list of arithmetic, logical, comparison, bitwise, assignment, and special operators; and, the functions that are used in the resolution process will be set as operators and their corresponding arguments as operands. For example, consider this code line written in python:

$$pks = peakdelta(s, delta = np.percentile(s, 70) - np.percentile(s, 30)) \tag{6.1}$$

- **Total List of Operators**: *'peakdelta'*, *'percentile'*, *'−'* and *'percentile'*
- **Total List of Operands**: *'s'*, *'delta=np.percentile(s, 70) - np.percentile(s, 30)'*, *'s'*, *'70'*, *'s'*, *'np.percentile(s,30)'*, *'np.percentile(s,70)'* and *'30'*
- **List of Distinct Operators**: *'peakdelta'*, *'percentile'* and *'−'*
- **List of Distinct Operands**: *'s'*, *'delta=np.percentile(s, 70) - np.percentile(s, 30)'*, *'np.percentile(s, 70)'*, *'np.percentile(s, 30)'*, *'70'* and *'30'*

In the second case, both **connotation** step and **search** procedure will be inspected for the **difficulty** measures. Regarding the **connotation** step, the evaluation will be made similarly to the previous method for inspecting functions, in which the function is set as an operator and its corresponding arguments as operands. For the **search** procedure, the set of instructions typically used in regex (except the "." character) is defined as operators[86]. These can be revisited in the second section. The remaining characters of the regex string, except the parenthesis, will be set as operands. In this are included the "." operator that matches any character, and the "\d" and "\w" that matches all digits and alphabetical characters respectively. Regarding the parenthesis, when occurring ("{}" or "[]"), will be set as one operator. For example:

$$Connotation : Sm\ 500\ A\ 0.8\ D1\ 0.05$$
$$Search : \texttt{(z1n).*?}$$

[3] https://github.com/novabiosignals/SyntacticSeachOnTimeSeries
[4] https://novabiosignals.github.io/SyntacticSearchonTimeSeries/

- **Total List of Operators**: *'Sm'*, *'A'*, *'D1'*, *'( )'* and *'*?'*
- **Total List of Operands**: *'500'*, *'0.8'*, *'0.05'*, *'z'*, *'1'*, *'n'* and *'.'*
- **List of Distinct Operators**: *'Sm'*, *'A'*, *'D1'*, *'( )'* and *'*?'*
- **List of Distinct Operands**: *'500'*, *'0.8'*, *'0.05'*, *'z'*, *'1'*, *'n'* and *'.'*

Table 6.6: Evaluation of the complexity in the resolution of examples with the *syntactic* approach and with the *classical* approach. Voc - Vocabulary, Lgth - Length, CL - Calculated Length, V - volume, D - difficulty, E - effort.

Example	Syntactic Resolution						Classical Resolution					
	Voc	***Lgth***	***CL***	***V***	***D***	***E***	***Voc***	***Lgth***	***CL***	***V***	***D***	***E***
Start/End of Plateau	5.0	5.0	8.0	11.6	1.5	**17.41**	12.0	13.0	33.3	46.6	2.5	**116.5**
Step Detection (Left & Right Heel)	5.0	6.0	8.0	13.9	1.75	**24.38**	19.0	22.0	63.6	93.5	4.2	**388.2**
Dicrotic Notch	7.0	7.0	13.6	19.7	2.0	**39.3**	21.0	25.0	73.0	109.8	4.7	**517.7**
Electrocardiogram Peak Detection	2.0	2.0	2.0	2.0	1.0	**2.0**	9.0	12.0	20.3	38.0	3.0	**114.0**
Straight Line Tracking	2.0	2.0	0.0	2.0	1.5	**3.0**	25.0	33.0	93.5	153.2	5.3	**811.3**
Stable Lifting Detection	10.0	18.0	24.4	59.8	3.6	**217.8**	22.0	23.0	77.3	102.6	5.0	**512.8**

Table 6.6 presents the performance measurements for both methods of solving the demonstrated examples. For the evaluation, the pre-processing step was omitted, which means that only the regex is used to compute the difficulty measures for the *syntactic* approach, and only the script after the pre-processing steps is used to compute the difficulty measures for the *classical* approach.

6.2.3 Discussion

The previous section demonstrated the potential of the proposed framework in delivering a more expressive methodology of solving query search tasks in time series. For the complete group of examples presented, all the steps involved were written in a symbolic manner. This fact eased the way in which time series are interpreted and how the final search is achieved.

The results from the *Halstead* measurements summarized in Table 6.6 demonstrate that, in general, the complexity is decreased using the SSTS approach in comparison with the classical solution. In the circumstances in which the difference is not significant, such as the *Example 6* (Stable Lifting Detection), the complexity measurements increase significantly. This fact suggests that for more complicated problems, the complexity of the regex increases.

6.2.3.1 Symbolic series generation

Overall, most of the symbolic series that are generated by the symbolic connotation step are simple to read and interpret. Nevertheless, the complexity of the series linearly increases with the number of operators used to perform the symbolic connotation. Consequently, for problems that have inherent and higher complexity, and require the use of several connotation groups, it is expected that the legibility of the symbolic sequence will also be reduced. *Example 6* constitutes an example of the mentioned drawback, as it involves two distinct resolution steps and revealed to be much more complicated to understand without any previous insights.

6.2.3.2 Required background knowledge

Each step of the SSTS pipeline has an important role in solving the examples, especially the two initial steps (pre-processing and symbolic connotation). Both steps aim to simplify the signal, such that the search procedure can be composed as easily as possible. This approach aims to prepare the signal so that a sequence of symbols can be used to simplify the search step. Therefore, choosing an adequate set of pre-processing methods and proper coordination with the symbolic connotation methods will reduce the search task complexity, both in the writing and interpretation processes.

SSTS users' should have background knowledge in both signal processing techniques and regex. In fact, without prior consolidation of these topics, the user might have an additional effort in generating a solution to solve the required problems. Firstly, for circumstances in which the user experiences difficulty in using such methods, the pre-processing and the symbolic connotation steps will not be so intuitive. Secondly, for in case the user is not familiarized with regex, it will be hard to generate appropriate expressions to search for patterns on the symbolic time series, especially for more difficult problems, which require the use of advanced operators.

As a future objective, we expect to increase the expressiveness of the tool and reduce the required knowledge to accomplish signal processing tasks and regex queries. In order to achieve a more simplified pipeline, a set of pre-processing and symbolic connotation templates for specific types of data (e.g. electrocardiogram) should be available to the user. Furthermore, the search step could be also simplified using meta-regex. In addition, a flexibility in the search process should be considered. regex are powerful tools, but lack flexibility, in the sense that the pattern is or not matched by the regex. Instead, a score function based on the query should be considered to make a flexible search and provide an error margin to the user.

We expect SSTS to have a broad range of applications that can extend to the query-based search problematic. Other topics in time series data mining, such as *signal generation*, *segmentation*, *feature extraction*, *labelling* and *classification*, could benefit from this reasoning.

In this thesis, only the classification topic was explored. In the next Section, we present our results.

6.3 Time Series Classification with HeaRTS

In this section, the performance of HeaRTS is evaluated in the time series classification task by being applied to the *UCR time series classification benchmark* and compared with the traditional 1-NN ED. As was mentioned on Chapter 5.2, this process can be performed in multiple ways, namely with the combination of a *Bayesian* or SVM classifiers with either a BoW or a TF-idf model. These combinations were also compared. This evaluation will help us answer if it is possible to use a text-based representation of time series with words extracted by SSTS queries with a traditional text-mining strategy.

In addition to these experiments, we performed an individual analysis of several use-cases to understand how to extract meaningful explanation about differences between signals based on this linguistic translation. This study should tell us if the words extracted are valuable to highlight the *subsequences* that are most relevant to distinguish the time series from one class in regard to all the others.

6.3.1 Classification Performance

Figure 5.9 showed us how vectorized documents can be compared with the cosine distance to distinguish time series. Simply by using part of the SSTS queries, it is possible to generate TF-idf weights that are differently distributed based on ordered words on a time series document. Of course, the *Trace* dataset used for this example is a simple dataset, with simple dynamics, but still challenging to classify well with the 1-NN ED (only 76% of accuracy). Therefore, to validate the ability of the proposed method to perform time series classification, it was computed on all available datasets of the UCR Benchmark and its performance compared with the standard 1-NN ED. In order to perform this evaluation, we followed the indications provided by the authors of the dataset [44]. As recommended, we used the reference *train/test* split, and performed a cross-validation for hyperparameter tuning with the standard *train* set.

As mentioned on Section 5.1, the method translates the time series into text with SSTS queries from Table 5.3. These have pre-processing, connotation, and search steps. For each step, relevant parameters are necessary depending on the queries used. For this experiment, we used as pre-processing a simple smoothing, which requires a *window size* (w_s). The derivative connotation steps also required a specific threshold (thr). In addition, the BoW and TF-idf models can be built with different *n-gram* sizes:

$$\begin{aligned} w_s &\in [1, 10, 25, 50, 100, 250] \\ thr &\in [0.001, 0.01, 0.05, 0.1] \\ n-gram &\in [1, 2, 3, 4, 5, 6] \end{aligned} \tag{6.2}$$

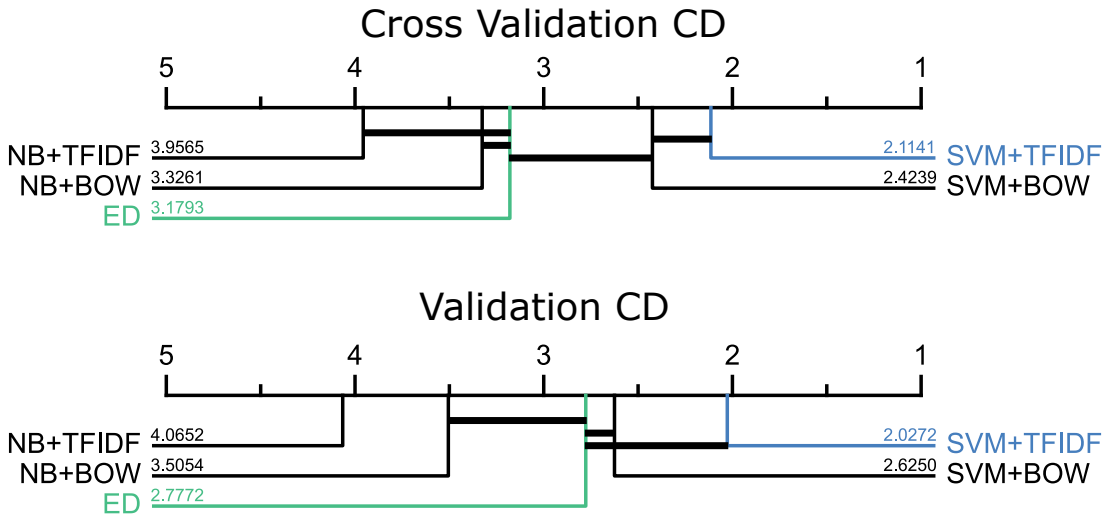


Figure 6.10: Critical distance diagram for the methods that rely in a textual representation of the time series (BoW and TF-idf) and euclidean distance for: **Top)** the cross-validation stage; and **Bottom)** the validation step. The closer to **1** is the method, the better it is. Thick lines between methods indicate that these have not a significant performance difference.

The cross-validation step for hyperparameter tuning was performed differently depending on the size of the dataset. For datasets with more than one hundred signals, we performed a 10 fold stratified (K-SF) cross validation, while on datasets with less than one hundred signals, a leave one out (LOO) cross validation was performed instead. The *K-SF* cross-validation step was chosen on larger datasets because of its higher computational speed performance in comparison with the *LOO*. In shorter datasets, using the *K-SF* would reduce the training set size too much and with unbalanced training samples for certain classes, and, as recommended by [44], we used the *LOO* instead. We then selected the parameters set with best average cross-validation F1-scores for each dataset and performed the validation on the *test* set, to which the model was *blind* to.

We performed this experiment with several strategies and combinations, namely: BoW with Bayesian Classifier (BoW+NB), BoW with SVM (BoW+SVM), TF-idf with Bayesian Classifier (TF-IDF+NB) and TF-idf with SVM. The summary of the performance is presented in Figure 6.10.

The performance was measured for both the cross-validation step and the validation step. The overall results can be found in Appendix B, Section B.4.2. As can be seen, the strategies that rely on a SVM classifier are more robust. The TF-idf turns out to not be reliable when used with the NB classifier, which has much better results when the BoW model is used instead. On the other end, the TF-idf model works well with the linear SVM classifier, being the method with highest average accuracy.

Figure 6.10 also shows that there is a significant distance in performance between the TFIDF+SVM model in comparison with the 1-NN ED method during the cross-validation step. However this significance is not as expressive in the validation step. All these remarks are also highlighted in Figure 6.11, where both models (BoW and TF-idf) and

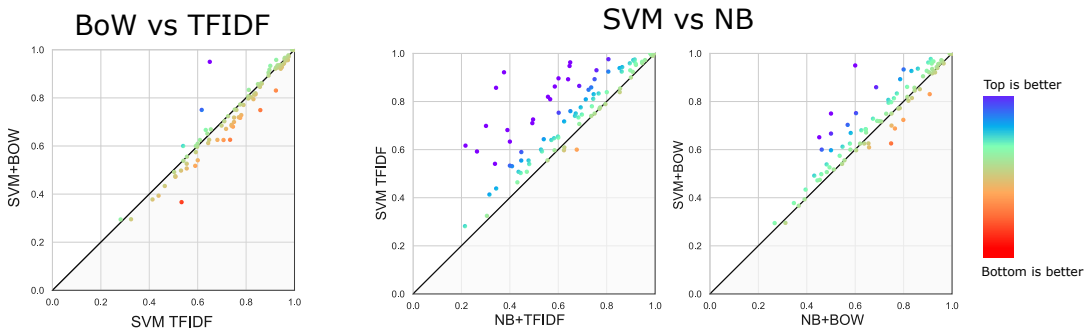


Figure 6.11: (left) Comparison of classification accuracies of HeaRTS when using the BoW or TF-idf. (right) Comparison of classification accuracies of HeaRTS when using an SVM or NB classifier. The colormap indicates the distance in accuracy between methods, being blue an indication of the upper left method being better, and red the bottom right method being better. Recall graph explanation from 2.8.4.1.

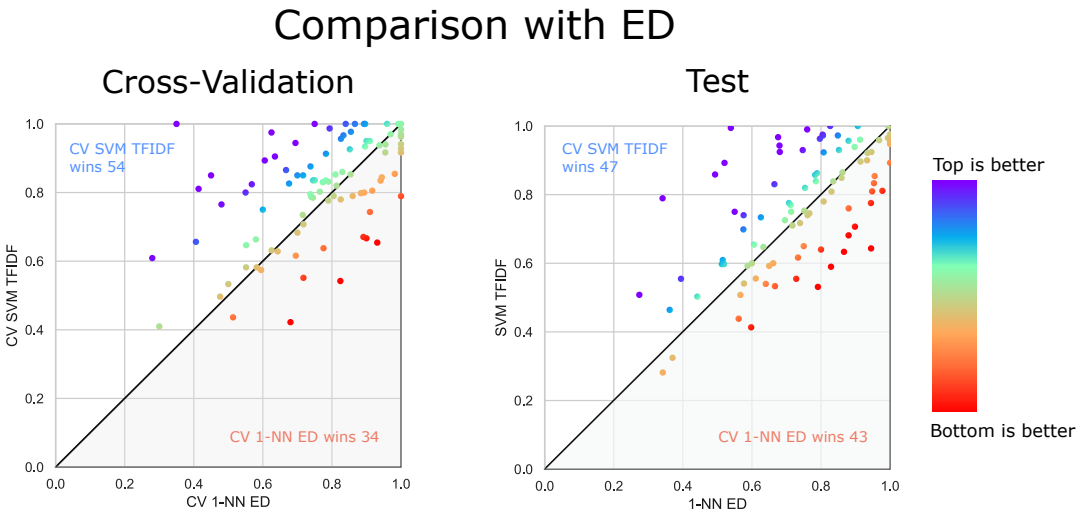


Figure 6.12: right) Classification accuracies for the Bag of Synthatic Patterns in comparison with the euclidean distance combined with 1-NN classifier (1-NN ED) with standard train and test; left) with a 10 fold cross validation. The colormap indicates the distance in accuracy between methods, being blue an indication of the upper left method being better, and red the bottom right method being better. Recall graph explanation from 2.8.4.1.

classifiers (SVM and NB) were compared (left and right, respectively). When comparing both models (BoW and TF-idf), we see that the best performances are similar, with a slightly better accuracy achieved with the TF-idf model. In regard to using the SVM or the NB, as mentioned previously, better results are achieve when using the SVM classifier, for both models (TFIDF on the left, and BOW on the right, of the right graphs).

We are now in position to compare the best text-based method (HeaRTS) with the 1-NN ED classifier. This result is presented in Figure 6.11, for both cross-validation and test steps.

The average accuracy for the total 93 datasets is 81%/76% for HeaRTS (with 54/47 wins) and 76%/74% for the 1-NN ED (with 34/44 wins). Analysing the results in more detail, we find that HeaRTS works poorly in problems with a high number of classes (e.g.

Phoneme, with 39 classes). Both *Haptics* and *InlineSkate* are badly classified in general by state of the art methods, and HeaRTS has similar results. Binary problems that rely in shapes well described with the derivative properties, such as *ShapeletSim*, *BirdChicken*, and *Trace*, are typically better classified. Other datasets, such as *GunPoint* or *Plane*, have a more comprehensible pattern extraction, with structures that can be better described with the SSTS queries used. The ability to perform the separation will be as good as the richness in the characterization of the classes and their differences. In datasets where the queries might identify a `peak` as something else, might affect negatively the ability of the classifier to correctly learn the dynamics of the signal.

We wanted to answer two main research questions: (1) can we solve time series classification tasks based on its textual representation following traditional *NLP* methods and (2) could we use the words that describe time series as a readable explanation of the data and the differences between classes. Having evidence that the first point is achieved, we will look into the second question.

6.3.2 Interpretable Data Outputs

Besides the fact that this approach can be used for classification tasks, we expected to go further by providing some feedback from the textual description, helping to understand what differentiates the classes. The TF-idf model gives weights for each *word* that characterizes the time series. By searching back the *words* on the time series and summing the weights corresponding to these *subsequences*, a shape relevance score can be computed, highlighting areas of higher interest that can be used to differentiate the different classes. The next Figures (6.13, 6.14 and 6.15) are displayed showing the highlighted areas based on the TF-idf weights.

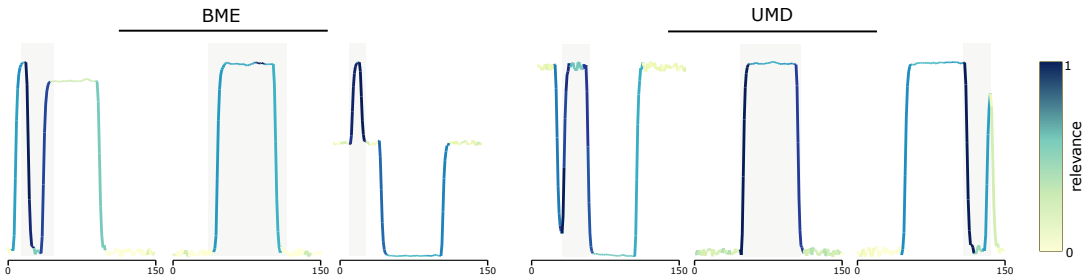


Figure 6.13: Interpretable results with highlighted shapes for the *UMD* and *BME* datasets. The *UMD* dataset was classified with with an accuracy of 100%, computed with a $w_s = 10$, $thr = 0.05$ and a $2 - gram$. The *BME* was classified with an accuracy of 100.0% with a $w_s = 10$, $thr = 0.05$ and a $2 - gram$

On Figure 6.13 are showed two similar datasets. The f1-score of this task was 1, meaning that it was able to perfectly characterize each shape with the textual representation. For instance, the *BME* dataset has two different sequences of shapes, since the first class has a `peak followed by a positive plateau`, while the third class has a `peak followed by a negative plateau`. The second class has simply a `plateau`. The

main difference between all those classes is the peaks and what follows them, which is precisely what is highlighted. The same happens on the *UMD* dataset, in which the most relevant element is the peak and the plateau. As we can see, the transition from the peak to the plateau is highlighted.

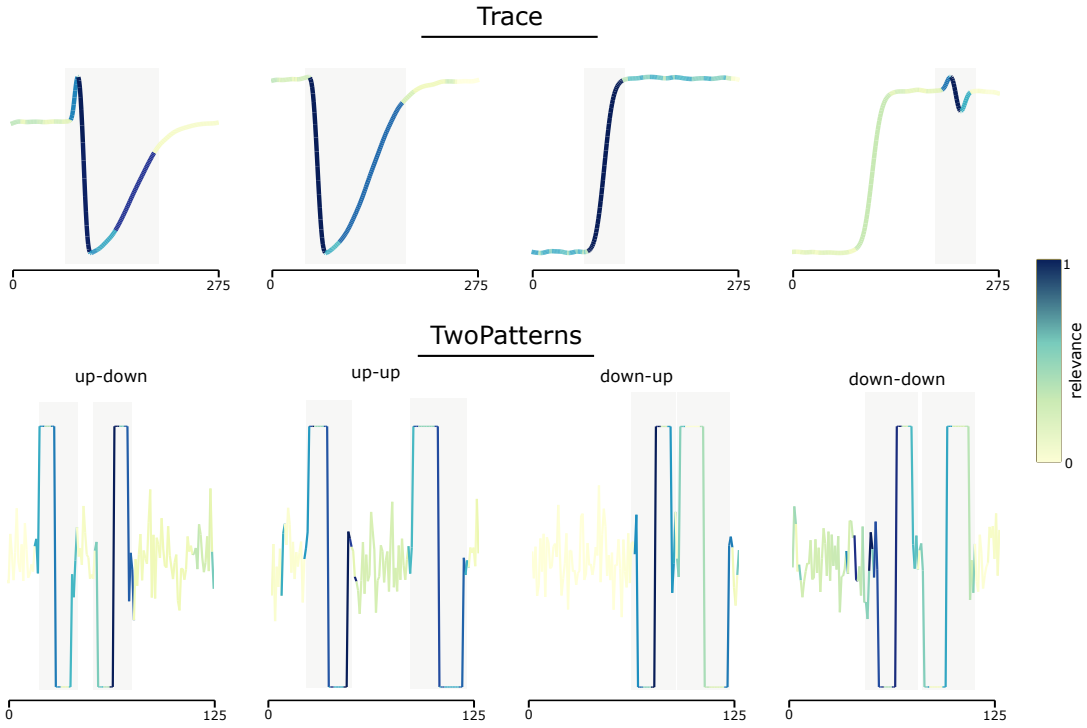


Figure 6.14: Interpretable results with highlighted shapes for the *Trace* and *TwoPatterns* datasets. The *Trace* dataset was classified with with an accuracy of 100.0%, computed with a $w_s = 25$, $thr = 0.05$ and a $1 - gram$. The *TwoPatterns* was classified with an accuracy of 100.0% with a $w_s = 10$, $thr = 0.1$ and a $5 - gram$

Another dataset that has intuitive differences between classes is the *Trace* dataset. The *subsequences* of interest are highlighted on each time series of each class. For instance, the first class has the `peak` and `valley` highlighted, contrasting with the `valley` from class 2. Classes 3 and 4 have notoriously different valued elements. While class 3 has the rising phase as the relevant shape, class 4 has a `small peak followed by a valley` on top highlighted.

The *TwoPatterns* dataset is also intuitive to understand and fits the problem for this experiment. Each class has mostly noise, filled with step functions that have different sequences. Class 1 starts by having an `up` step function and then a `down` step function. The same logic is applied to the other classes. The proposed method highlights these step functions as being the relevant segments of the signal.

Finally, two other examples are also presented. One of the *GunPoint* datasets and another from the *ShapeletSim* dataset. The *GunPoint* has 2 classes: (1) the subject draws a plastic gun and points it; (2) the subject simply points with his/her hand. The major difference between the two classes is the drawing process, highlighted by the method as

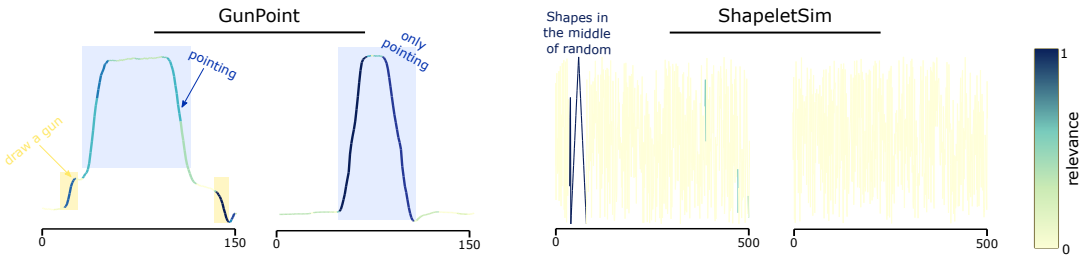


Figure 6.15: Interpretable results with highlighted shapes for the *Gunpoint* and *ShapeletSim* datasets. The GunPoint dataset was classified with with an accuracy of 96.0%, computed with a $w_s = 10$, $thr = 0.05$ and a $2-gram$. The *ShapeletSim* was classified with an accuracy of 99.0% with a $w_s = 1$, $thr = 0.05$ and a $1-gram$

the `small rise` and `small fall`.

Regarding the *ShapeletSim* dataset, the data has also two classes. In one, simple shapes, such as triangular, square, or sawtooth waves were added to a random signal. The method can highlight the representative shape of the signal.

6.3.3 Discussion

In this section, we wanted to answer two main research questions: (1) can we solve time series classification tasks based on their textual representation following traditional *NLP* methods, and (2) could we use the words that describe the time series to hint about and explain the differences between classes of time series. For this purpose, we applied the proposed method on the *UCR* classification benchmark and highlighted several relevant use-cases and the overall results achieved.

The results provide evidence that it is possible to compare time series based on a textual representation and use traditional text-mining strategies for classification purposes. Not only is it possible to compare the vectors of time series documents with the cosine distance (as we have seen on Figure 5.9, but BoW or TF-idf models can be used with a linear SVM to solve classification tasks.

Additionally, the fact that words are weighted based on their relevance to distinguish each time series document from all the others, it is possible to highlight the *subsequences* of the time series corresponding to these words and understand which parts of the signal are more relevant, being a step forward into interpretability of time series. The limited examples were still simple and intuitive and the process turns harder to understand in more complex data. In these cases, it is hard to truly find meaningful differences and explain why these classes are different, especially with words. This limits the further interpretability potential of this method because the differences might not be perceptible by the naked eye. On one end, the performance can be improved using a richer and correct set of textual descriptors. We mean richer because the differentiation is as good as the description performed, and we mean correct because the translation process from the numerical domain to the text domain is not error-proof, which means that a mistranslation may occur and make the results worse. This fact can also contribute to the lack in ability

to generalize the model from the cross-validation to the validation steps, with an evident decrease in performance.

Another relevant point is the fact that words used might not be capable of expressing the differences between classes. Many cases with high accuracies can be highlighted, but their interpretability falls short. For instance, for the dataset *ECGFiveDays*, a binary classification problem. The method has 0.96 of accuracy, but does not show expressively the difference between classes.

The textual representation is valuable in the sense that text mining domain knowledge can be used on time series. Nevertheless, we can profit from this knowledge as much as the translation of the time series into text is rich, valuable, and domain-specific. With the current set of SSTS queries we can make a limited description that works well in simple signals described mostly by their derivative dynamics. The keywords extracted for these time series still have limited readability being even harder for more complex examples. However, if more expressive queries are used to make a more robust translation, the keywords extracted can be more valuable and relatable to the time series studied. Therefore, a more diverse set of queries should be used to make possible the extraction of more relevant keywords. At some point, a reduction and selection of the most relevant keywords should be made.

6.4 Pattern Search with QuoTS

In this Section, we demonstrate the potential utility of QuoTS to search for relevant subsequences in real datasets. Table 5.4 from the previous Chapter highlights the keywords and operators if needed to follow these examples. This method, in contrast with SSTS, uses features and not a symbolic representation of the time series.

In order to show evidence that QuoTS is useful in a multitude of scenarios, we present several examples of its application to find relevant patterns on real domain time series. The section will start by presenting how well it matches hand gestures; then show its applicability in searching for relevant patterns in several problems and use-cases; and finally, we present the ability to include additional words based on existing patterns, with the example of *puppeteering* a toy car model.

6.4.1 *QuoTS* Matches Gestures

We start by demonstrating the ability of QuoTS to sort how well individual shapes match a written query. For this, we used the *UWaveGestureLibrary* dataset from the *UCRArchive* as a proxy [110], which similar to telemetry data, relies on inertial time series. We use this proxy data simply because it is much larger than any publicly available labeled transportation data. However, we note that gestural interaction with the automobile interfaces is an area of active research [40, 176]. With this example, we show that the system can recognize

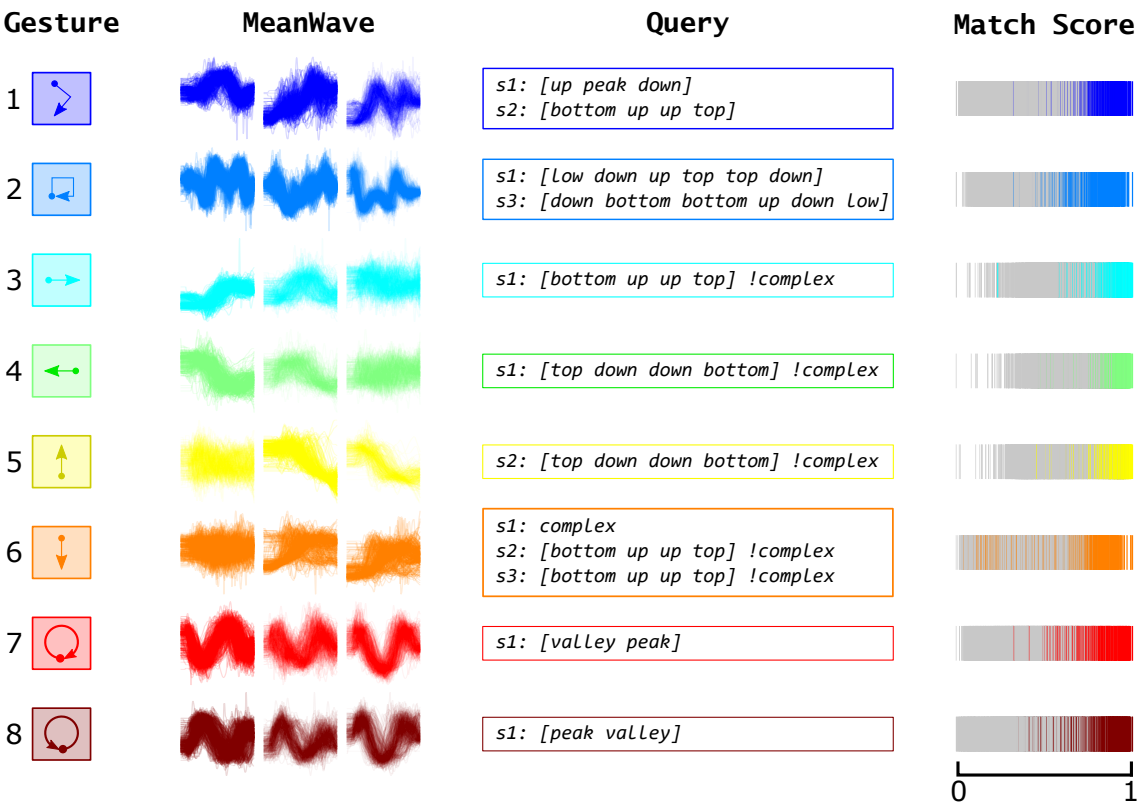


Figure 6.16: *UWaveGestureLibrary* subsequence matches with QuoTS. Column-wise are showed the gesture class, the corresponding mean wave for all subsequences, the query used to match the subsequence class and the corresponding match score for all subsequences, highlighting with the color of the specific class.

different gestures with or without intuitive queries, hence, if humans were able to learn and master this tool, this recognition problem would be largely solved.

We not only demonstrate that the system can significantly distinguish gesture patterns, but it does so with very simple and intuitive queries. The ability of the written queries to match the correct class is presented both visually (Figure 6.16 and quantitatively (Table 6.7). In Figure 5, the set of eight gestures is described row-wise, having a corresponding mean shape (**MeanWave**) and a query (**Query**) that should filter it. The visual intuition is demonstrated with the barplot (**MatchScore**), which for the sake of readability, highlights the normalized match score ([0-1]) of subsequences belonging to the row-wise specific gesture. The other classes are shown in gray. For each gesture, we show the shape of all the available axis (X, Y, and Z). We attempt to create queries using only a single axis (X-axis) but used other dimensions when needed.

To quantify these results, we looked at the top-10 and top-100 matches for each class and counted how many gestures of the selected class were correctly sorted (TP/10 and TP/100). The results are presented in Table 6.7. These show that the top 10 matches always correspond to the correct class. The reader might think that the first 10 matches might be too easy considering a problem that has around 400 gesture samples per class, but note that having more gesture samples also means more opportunities to make mistakes. Moreover,

	G1	G2	G3	G4	G5	G6	G7	G8
TP/10	10	10	10	10	10	10	10	10
TP/100	96	100	94	95	74	100	100	99

Table 6.7: Results for the top 10 and top 100 sorted gestures classes (G) when using the queries from Figure 6.16.

considering the analogy of searching for a webpage with the *Google* search engine, a user will probably be interested in examining at least the top 10 results. Nevertheless, the reader will appreciate that even if we consider the top 100 matches, QuoTS still achieved impressive results.

Although we simply wanted to demonstrate that with a set of meaningful words we can correctly sort each of the classes of this dataset, we also want to highlight that the nature of the classes can be very well expressed by the queries. This is especially evident when we look at the query for gestures that are inverse to each other, such as gestures 7 and 8. Gesture 7 is well matched by the query `[valley peak]`, implicitly, the transposed gesture should have the exact opposite query, which it indeed does (`[peak valley]`). This also occurs for the two other sets of gestures that occur in natural pair; gestures 3 (`[bottom up up top] simple`) - 4 (`[top down down bottom] simple`) and gestures 5 (`[top down down bottom]`) - 6 (`[bottom up up top]`). We also demonstrate that for the handful of cases where this did not work, there was a semantic explanation for it. (e.g. Classes 5 and 6 have not a specific pattern and seem to be especially random in their X-axis, or classes 1 and 2 are very similar, and because of that, are mislabeled). But, as our tool can perform queries in multidimensional data, we can still discriminate these by using the Y-axis in conjunction with X-axis to sort them correctly. Note that discriminating among these eight classes is not a trivial problem. Of the more than 1,000 papers to have worked with this dataset, the current best accuracy was obtained by *COTE* algorithm which achieved 76.56% using a single axis. In addition, this dataset was acquired from eight different subjects, which indicates that our system can account for the intra-subject and inter-subject variability in motion for this dataset [110].

6.4.2 *QuoTS* in Selected Use Cases

In this section, we present several examples of where QuoTS can be used in continuous real data. We start with an ECG signal that has motion artifacts.

6.4.2.1 Pattern search on the ECG

As a start, we were searching for the motion artifact segment that appears in two areas of the signal. This artifact can be detected in two ways, as we are showing in Figure 6.17. The first would be to use the keyword `noise`, which searches for parts of the signal that have a higher difference from a moving average of the signal. In this case, the higher difference

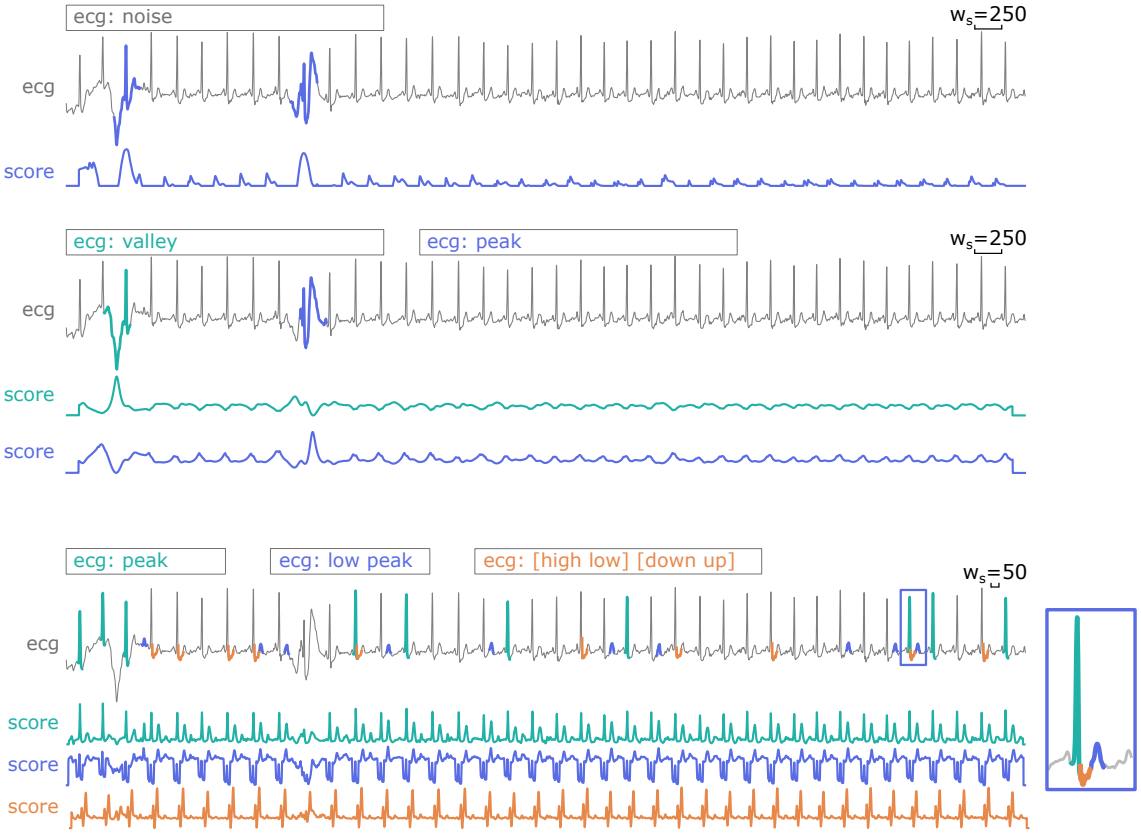


Figure 6.17: ECG use case to identify noisy sections, as well as specific segments of the ECG pattern. The queries are written in text boxes. On the side, an example of a match is shown as a larger pattern.

occurs when the motion artifacts appear. We could also search for each of these motion artifact sequences simply by pure visual intuition of their shape. For instance, the first motion artifact looks like a `valley`, while the second looks like a `peak`, which ends up matching the desired segments. This process was made by looking for subsequences with 250 samples.

On the ECG signal, we can also find other segments of interest, namely from the *PQRS* complex. In this case, we searched for the *R* `peak`, the *S* `valley` and the *T* `peak`. The first subsequence was simply matched by searching for the keyword `peak`, with a window size of 50 samples, which contrasts with the 250 samples from the previous example. The keyword matched well the highest peak, but other peaks were present with lower amplitudes, such as the *T* `peak`. This one was matched with the query `low peak`. In between these two subsequences, there is the *S* `valley`. In this case, we show the usage of the special `followed by` operator. The query `[high low] [down up]` indicates that the first half of the searched subsequence should have a *high change in amplitude going down* and ends with a *low change in amplitude going up*. This is exactly what is matched, as highlighted and indicated by the corresponding *score* function. To be clear, the colored highlighted segments represent the 10-most relevant matches for the written query. The *scoring* functions show that the other matches would be found as well.

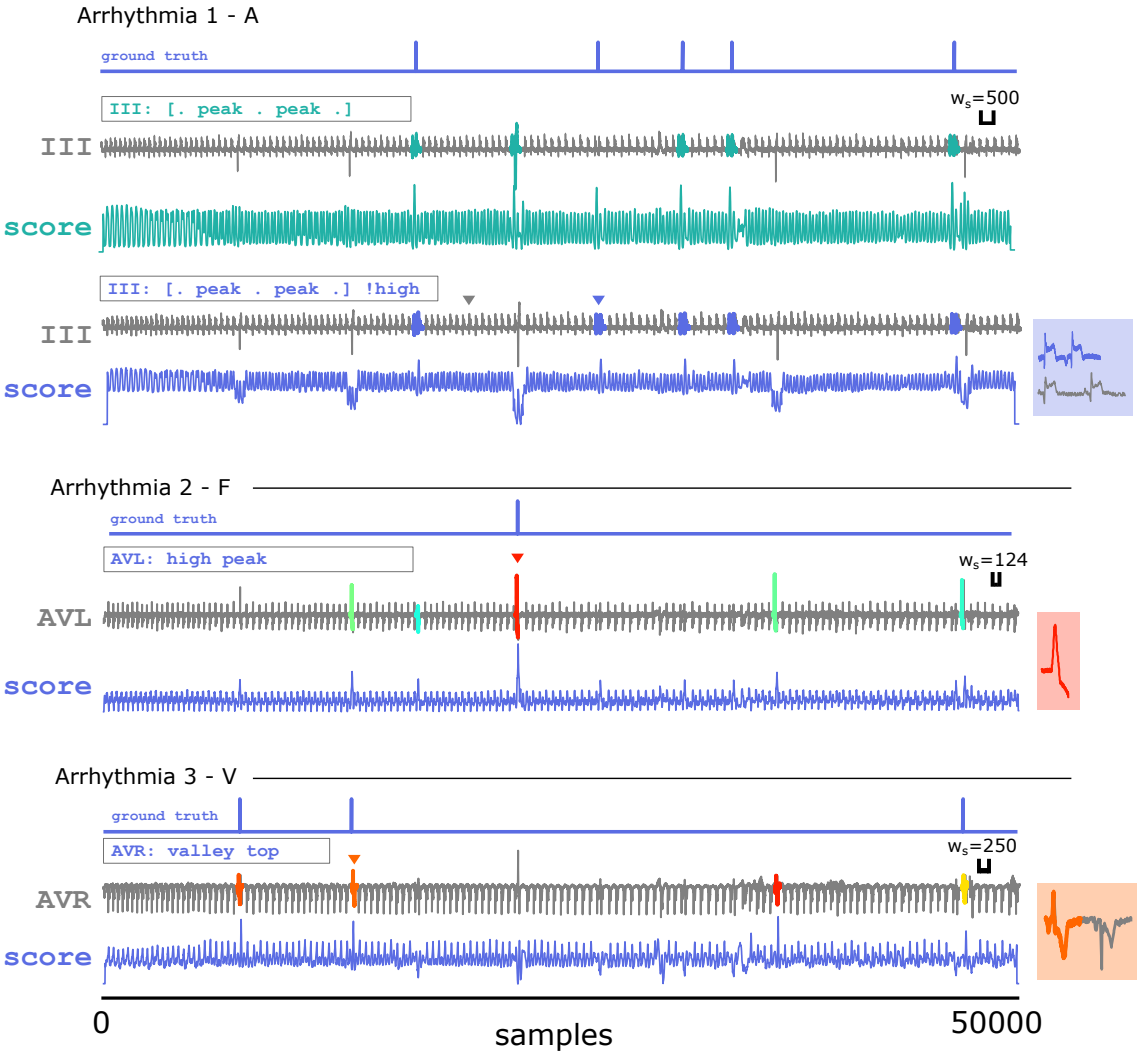


Figure 6.18: Examples of the detection of three specific cases of arrhythmia. The ground truth is selected from the annotations of specialists in the area. The queries are presented in text boxes and the found patterns are highlighted. On the side, an example of a match in a larger size is shown.

The ECG signal is a perfect example of how a text-based search can help find relevant occurrences, as we showed previously. However, there are occurrence of higher interest, such as patterns linked to specific medical conditions that are represented by variations in the original shape of the ECG signal. Figure (6.18) shows several examples of the detection of three different types of arrhythmia from the *St Petersburg INCART 12-lead Arrhythmia Database* (Physionet) [60].

The ground truth follows the annotations present on the original files. Three types of annotations were found, associated with different types of arrhythmia events (A -  , F - and V -). Using QuoTS, we can search for these occurrences in the presented segment. The first event is characterized by having a normal beat irregularly before it was supposed to. It is then possible to find two ECG beats in a shorter window. This means that if we search for two main peaks inside a short window, separated by anything (`.`) `III:`

`[. peak . peak .]` (in this case 500 samples on the lead III), it should be possible to find where these events occur. We displayed the 5-top subsequences that match the query and the corresponding scoring function. Four out of the five events are highlighted, but the scoring function indicates that the "missing" subsequence has a high match value as well. The wrongly identified event fits the description as well, being a different type of arrhythmia, but with a different shape (type F). This shape is normally a high change in amplitude, therefore if we state that we reject any high amplitude change subsequence, we should be able to match the five desired subsequences. Adding `!high` to the previous query (`III: [. peak . peak .] !high`) leads to correctly matching all the desired subsequences. To be clear, we chose lead III without any specific purpose, being possible to use the other leads for the same purpose. What matters is that the desired subsequence has a particular characteristic that differentiates it from the rest.

The second example corresponds to the arrhythmia of type F, characterized by a high change in the amplitude of the ECG with an irregular beat. In this case, we can match this event with the simple query `AVL: [high peak]`. Other events are matched but have a much lower amplitude. The other events matched are from the type V, which does not present a long flat section after the irregular beat. In this case, the lead AVR shows a very specific difference from the other two types of arrhythmia, which is the fact that the beat occurs on top, deviating from the middle of the signal. The shape continues to be a `valley` but on `top` (`AVR: valley top`). It is relevant to point out that the original annotations (ground truth) do not include one of the highlighted subsequences. Although not confirmed with a specialist, the shape of the event in all ECG leads are the same for this type of arrhythmia, so we suspect that there was a missing label that the query was able to indicate.

6.4.2.2 Pattern search on multimodal signals

In addition to searching for specific shapes on signals, QuoTS can be very useful for exploratory purposes in multivariate datasets. For instance, consider the following examples on a dataset acquired with the purpose to identify workload on drivers [173]. In this experiment, drivers physiological signals, such as ECG and Skin Conductance Response (SCR), were monitored. Exploring this dataset with QuoTS led to discovering several interesting occurrences, as shown in Figure 6.19. In this Figure are signaled specific events with road signs. We matched these signs' positions by finding their *latitude* and *longitude* on the map and matching them with the instant in time these coordinates occur on the dataset. In addition to this information, we recall the common terms used for auto telemetry, namely: Surge - breaking (peak) and accelerating (valley); and Sway - turning right (peak) or left (valley).

The first found events were identified by searching for moments where the driver was pulling the breaks and at a similar moment in time, the driver was having an increasing value of skin conductance (indicative of higher sweat and possibly momentary stress).

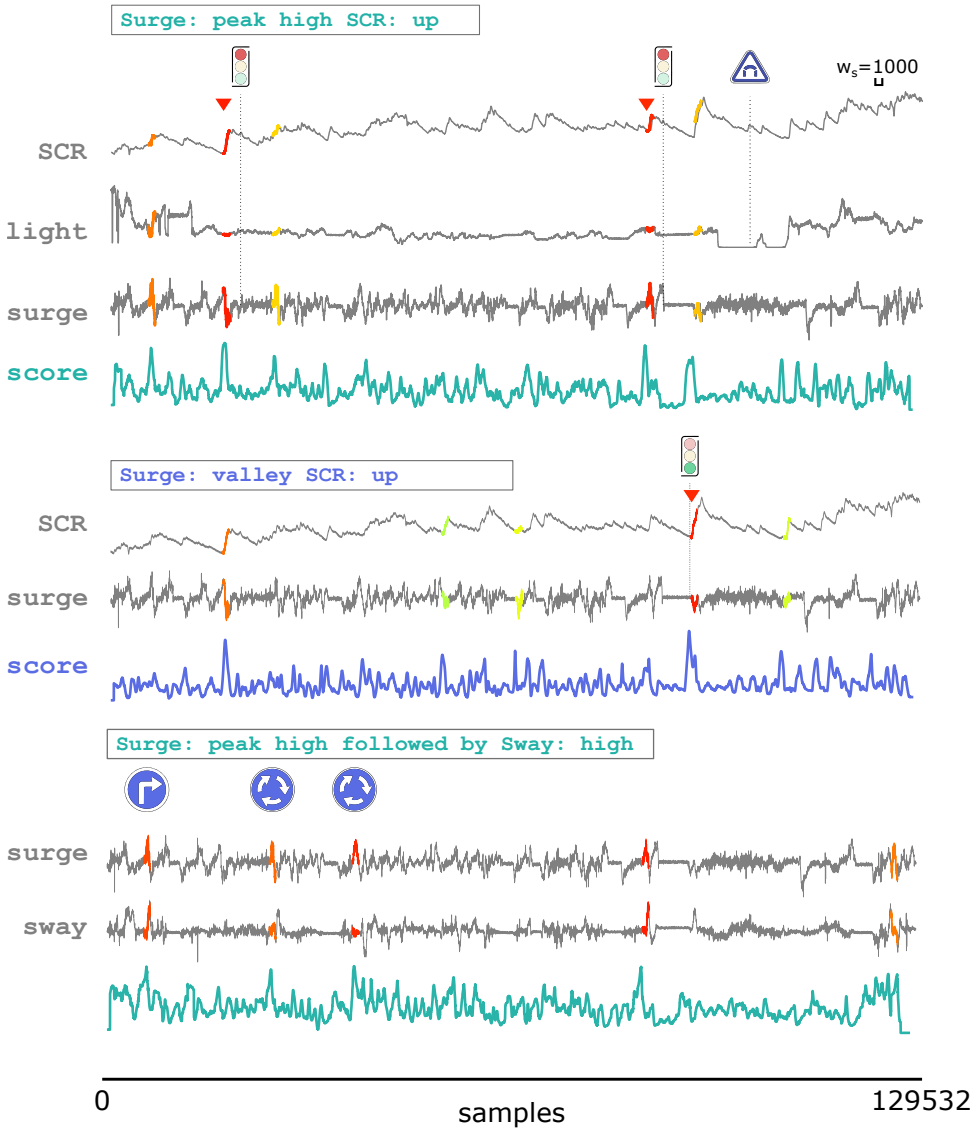


Figure 6.19: Example of multivariate patterns on a dataset from auto telemetry with physiological signals of the driver. The queries are written in text boxes. Several matches are highlighted based on the queries written. These events are associated with specific occurrences during the driving session, symbolized by traffic signs. These events were matched by searching the coordinates on the map. [173]

The query can be written as `Surge: peak high SCR: up`. We found that the presence of such events is very high on two occasions, in both cases, right before a traffic light. The second traffic light was turned red because the car stopped. This led to searching for a moment where the driver started accelerating and the skin conductance was increasing (`Surge: valley SCR: up`), which points to two main locations on the signal. The second is exactly when the car started driving after stopping at the traffic light, suggesting to be a relevant reaction from the driver. We did not have access to the videos, and can only guess what happened during the driving experience, but we can conclude that these identified moments are relevant to be observed afterward on the video, highlighting

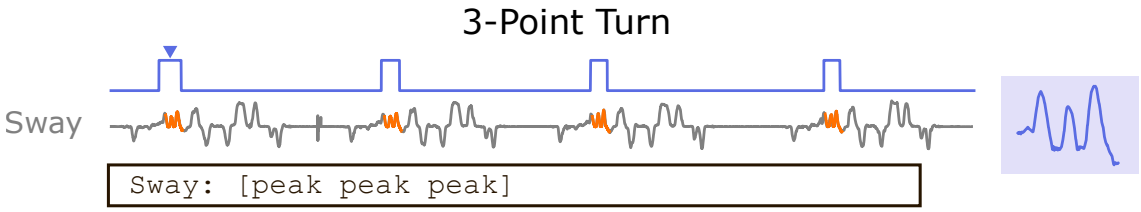


Figure 6.20: Example of searching for a 3 point turn event with QuoTS.

that QuoTS can be used to explore the signals very quickly, identify moments of interest (especially correlating multivariate variables) and search them on the video to validate what happened.

In addition to these events, we searched for moments where the driver had to break and then turned. There is a time dependency between these two occurrences and can use the `followed by` operator for this purpose. Therefore, we searched for `Surge: peak high followed by Sway: high`. The results pointed out several locations on the signal, three of them associated with turning right at a crossing and two roundabouts.

6.4.3 Matching *Known* Shapes with Words

We believe that QuoTS can be developed to be intuitive enough to allow most novice users to create simple effective queries for most information retrieval tasks. However, in some cases, the user's ability to formulate queries may be a limitation. In this section, we show that it is possible to perform a *mimicry* of the process being searched and add the query designed with the mimicry process as a *word feature vector* inside our vocabulary. We show a few examples of possible applications, starting with "puppeteering" a model car in performing a 3-point-turn. First, we will show the signal where we searched this event and how we used QuoTS to find it with a text query.

Figure 6.20 shows the search for a 3-point turn on a signal acquired with a smartphone inside a car while performing several driving exercises in a parking lot. Details of the experiment can be found at the Github repository [5]. The query used to match this event is very intuitive, being `[peak peak peak]`, which is exactly what this pattern is ().

Having found the desired pattern with text, we thought that it would also be possible to find it using a specific pattern and give it a name to use on QuoTS. In this case, the pattern was acquired with a toy model car, on which a smartphone was attached (as indicated in Figure 6.21) and acquired inertial data while mimicking a 3-point turn event on a table. The resulting shape is illustrated in *puppet data*. We believe this can be used in two ways: for the user to gain intuition as to how a motion/*manoeuver* is illustrated on motion data or how the inertial data from the motion/*manoeuver* can be used as a template on QuoTS:

• **Shape Intuition:** A user might not know what the shape of a 3-point turn looks like. By using a model car, he/she can mimic the motion and gain intuition over what the query

[5] https://github.com/Anonymous14151/QuoTS

should be (in this case, `[peak peak peak]`).

• **MASS template:** As mentioned above, we have a special shape word, which corresponds to a shape template given by the user, to which a word can be assigned. We then can use the word in our language to match desired patterns with the MASS distance profile as a word feature vector.

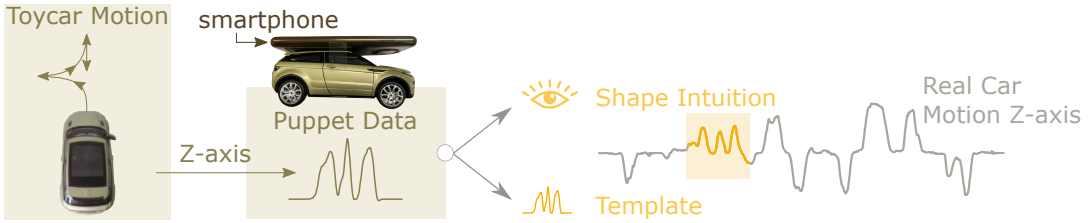


Figure 6.21: Creating query prompt data by puppeteering a car model. The puppet data for a 3-point turn (smoothed with 10 samples moving average) is presented on the left, while the real data from a real car is presented on the right.

We added the puppet data and the word `3pointturn` to our vocabulary. Searching for it in the signal matched all the desired shapes. Note that the template has a different amplitude and time scale considering that the motion was made with our hand, but the z-normalization makes the amplitude difference irrelevant, and the time scale was adjusted by resampling the template based on the selected window size for the query search. With this, the template is not only a template, it also becomes part of the language, being used in combination with all the other words available in our vocabulary.

7

Conclusion and Future Work

In this work, fundamental topics in time series data mining were researched. These topics were motivation by considering the need for tools that help analysts to better understand what happened during the recording process and find relevant instances on signals that can be related with specific occurrences in the physical world. These motivations are intrinsically related with tools that are more visually interpretable and search mechanisms that are more expressive and intuitive. The work developed contributed to each of these domains with relevant standard mechanisms that can be further developed into practical tools for real scenarios. Not only these tool can be practical, but the methods presented bring novelty into the state of the art, either by "borrowing" more traditional methods from other domains (such as the SSM in audio information retrieval or Natural Language Processing (NLP) techniques for text mining) or by introducing novel concepts, specially in terms of text representation of biosignals and text-based query search mechanisms.

In this Chapter, we highlight the main contributions of this thesis in each of this domains. Comments are also given regarding the contributions and applicability to occupational health scenarios. Finally, overall scientific production and collaborations during the period of this thesis are also provided.

7.1 Main Contributions on General Topics

As we mentioned in Chapter 1, this thesis contributes to three main topics, namely *Sensing*, *Analysis* and *Decision Making*:

- *Sensing:* In this context, a deep understanding of the existing technology to record biosignals was made. At some point, the focus moved towards technology existing to monitor occupational variables. A written work is available regarding the usage of fiber-optics for the monitoring of motion and postural variables in automotive industries, which can be found in [155] and it is compared to other existing methods. In addition, a study of the state of the art sensors existing on the market was made to understand the fit in the occupational health problematic. A presentation is available at [156], showing the existing materials on the market and that cover human motion

and physiology, more focused for occupational health. Several publications were also made regarding the setup and usage of inertial sensors in occupational settings [167, 166, 139, 161, 180]. The contributions in this area are more related with the knowledge gained and how it was used to more appropriately prepare an acquisition plan regarding the acquisition of occupational variables in several settings, namely automotive industry (Volkswagen Autoeuropa), clothing manufacturers and office/desk jobs.

- *Analysis:* This work has demonstrated a deeper exploration of the topics related with the analysis of biosignals and time series in general. The contributions include the usage of the SSM for the segmentation of time series (*novelty* and *periodic* functions), summarization and creation of similarity profiles for (semi-)automatic clustering. A novel symbolic representation of time series was introduced, and examples were provided in how to apply it for pattern search with regex (SSTS) and text-based classification (HeaRTS). A novel search mechanism was also developed, closer to natural language search, with keywords and operators (QuoTS). A more detailed explanation of these contributions are provided on a further section.

- *Decision Making:* The main purpose of the studied and developed methods was to help analysts when inspecting time series, but also move towards democratizing pattern search on time series. The proposed strategies provide several levels of understanding and help in several layers of decision making. Of course it will depend on the purpose, context, output delivered by the method and the category of expertise of the analyst.

  Starting with the analysis of structural information with the SSM, it provides a visual output that has characteristic structures that will appear independently of the type of data or context. The main information available on the SSM will always be *blocks*, *paths* and *similarity*. As we have seen, these can help the analyst identify specific occurrences on time series, with special interest in biosignals applications. Therefore, by learning how to retrieve information from the SSM, more awareness is given to the analyst and more conscious decisions can be made based on that information.

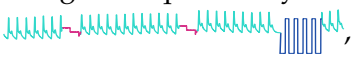
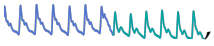
  From a standard perspective, the analyst can also perform automatic or semi-automatic segmentation and labeling/annotation of data to accelerate the preparation process to train supervised machine learning algorithms. This can either be done with the visual support of the SSM or the methods developed to extract the information automatically.

  Regarding the process of pattern search with more expressive queries, we believe the tools provide valuable resources to both the analyst that works in the time series

domain and is experienced in the computer science field (e.g. a researcher that develops machine learning methods for biosignals processing), and the analyst that is not an expert in computer science but has knowledge in a specific domain of time series (e.g. a physician that is experienced in ECG data). In this case, the process is more interactive and requires writing a regex/text query. This should be useful to accelerate the search process by an experience computer scientist, because it should be more quick to make the search with this system than developing an algorithm for that purpose. In addition, it contributes to democratize the search mechanism to more analysts than only computer scientists, because if the analyst is able to describe the shape being searched, it should be possible to find it (considering the right connotation methods/word feature vectors are used). In that aspect, the information retrieval would quickly help the analyst in taking more informed decisions.

7.2 Scientific Contributions

7.2.1 Unveiling the *Grammar* of Time Series

One of the major topics of this work regarded the segmentation of time series into smaller segments, based on *novelty* and *periodicity*. In addition, it was also discussed the benefit of relating the resulting subsequences by how similar these are. As an example, we showed the ABP signal  , which can be divided into 7 segments, having the structure A B A B A C A. We demonstrated with strong evidences that the usage of the SSM is reliable in performing this type of task. From the SSM, the novelty function can be extracted and the similarity profiles can be compared to perform a segmentation and association between subsequences. In addition, the segmentation might be periodic, meaning that a signal, such as  , can be separated into A and B, but also, AAAAAAABBBBBB. We also demonstrated that using the SSM, we can compute the similarity function, which highlights the cyclic nature of the subsequence.

The performance of the method was validated for the novelty segmentation process. It was compared to several SOA methods, showing to be competitive in tasks related with change point detection and segmentation. In addition, several use-cases from various fields were presented as examples, showing that the algorithm is agnostic to the type of signal, applicable to multidimensional and multimodal signals, and not requiring any previous knowledge on the data, such as the number of segmentation points. It also shows potential to be used for unsupervised annotation of data, being developed towards this purpose.

In this work, datasets were chosen to cover as much real scenarios as possible. Also, common benchmarks were used to validate the methods to guarantee independence from private datasets. Besides, the proposed method was also demonstrated to work well with multimodal and multidimensional occupational data.

There are still several improvements to be made, namely considering the excessive memory that is required to compute the SSM in cases where the signal is very large. The fact that a matrix has to be computed limits the ability to analyze in one run the entire signal. This process is specially relevant for periodic segmentation and computing similarity profiles. As previously explained, for novelty segmentation, the process can be adapted to only compute the SSM along the diagonal with the size of the kernel width. An additional limitation is the fact that the method is not invariant to trend. If events occur in slow trend changes, that is, a continuous linear change, the method will have more difficulty in identifying the segmentation point. The usage of pre-processing, additional time series representations, or time series decomposition methods could help in counteracting this effect.

The ability of the method to be adapted in Online scenarios has not been discussed, but a solution should also be considered in the future.

7.2.2 Using Language for Time Series Data Mining

Representing data into different data types provides a new look on the original data. It may lead to find segments of interest that were not visible in the original data type and/or may benefit from the large experience in mining on this new data type. These are the first arguments for the ideas we presented in this work in transforming time series from the numerical domain to the text/symbolic domain. A new look on time series is possible, which enables to adapt some of the existing data mining techniques with a textual approach, and it can benefit from the large knowledge on text processing or NLP.

As a first concept, language and time series can apparently be a strange combination. However, we provided additional evidence that there is a bridge and a potential to perform several tasks with success, namely in text-based query pattern search with SSTS and QuoTS, as well as classification with HeaRTS.

The concept of symbolic representation has started with SAX [103], which was a great inspiration for the work developed further with SSTS. We developed this method with the purpose of making pattern search with more expressive queries, more closely related with the way we look and interpret visually the existing shapes on time series. Several examples were provided showing the possibility of using regex (text patterns) on time series symbolic representation. Additionally, having higher levels of representation could be useful to search for increasingly higher-leveled structures, such as `peaks`, `plateaus` or a combination of these. This idea was what led us perform the next method for time series classification, HeaRTS.

We imagined that if time series could be *translated* into text documents, these could be differentiated based on the words and sentences that would represent them. Inspired by NLP techniques used for this purpose, such as BoW and TF-idf, we performed classification of time series documents, by creating a high-leveled distance measure that relies in the presence of structures, such as `peak`, `plateau`, `up`, `down` and `flat`, and the order

of these words in a sentence with *ngrams*. We then showed that it was possible to use traditional NLP methods to perform this task on the UCR classification benchmark. It was able to have a better performance than the 1-NN ED and showed to be competitive in this field. Especially because there is the possibility of extracting valuable information from the textual translation, namely by using the TF-idf weights to highlight areas of relevance on the signal or keywords that mostly represent the topic of the signal (such as topic modeling on text domain). We believe we introduced a novel idea with these processes that can be helpful for search but also for explaining which are the differences between signals. There is still a lot to improve, since as we showed, it would only be interpretable for time series with simple characteristics explained by the *connotation* and *queries* developed and used to describe the signals.

Having queries closer to the way we express what we see was one of the main motivations of this work. This led to the development of SSTS in the symbolic domain, but we believe as well that features are a good match to specific words that we used to describe parts of signals. This led to the development of QuoTS, which uses word-feature vectors to search for specific subsequences on time series by how well these match the set of keywords used, similarly to how we type keywords on *Google* to search for web pages. We provided evidence of its usage in several types of signal and with several types of problems, from motion gestures, to ECG patterns or telemetry data, in multidimensional time series. We highlight the potential to use this method to search subsequences based on visual intuition but also for *words* that are *known*, namely by *puppeteering* or *mimicking* shapes in practice. This is specially relevant if keywords can be transformed for each domain, being domain specific. Considering that the vocabulary can change from domain to domain, for instance, `peak` in medicine can mean an ECG peak, while in automotive telemetry, it might mean *turn right*. This domain specific match can help other non-experienced analysts to use it to search for specific patterns.

7.3 Other Contributions

7.3.1 Managing Rotation Plans with Exposure, Diversity and Team Homogeneity

In close collaboration with Volkswagen Autoeuropa and the Faculty of Human Motricity of Lisbon (FMH), we developed a method to automatically suggest job rotation schedules based on ergonomic standards available at the factory. These standard factors are from the AutoErgo tool, based on EAWS measures. The motivation for the development of such a tool was to help team leaders to manage job rotation schedules more quickly and in a more informed way. Team leaders organize the working schedule for their team by assigning each worker to a sequence of workstations for the entire week, which is a time consuming tasks and not always informed in the risk level that each tasks represents for a worker. In this method, risk exposure, diversity in exposure, as well as team homogeneity,

are taken into consideration when suggesting a daily rotation plan. The process was made by developing a genetic based optimization algorithm that followed an objective function developed by our team. The motivation, algorithm and results can be seen at [12].

This work was developed at the initial stage of the PhD, being valuable to get a better intuition of biosignals related with human motion. Having gained this intuition helped in understanding relevant aspects of signals and transfer this gained knowledge into the algorithms developed. For instance, problems such as segmentation were common in the signals acquired, which promoted the development of methods that could solve them (SSM). In addition, the shape of signals, associated with specific motions, were interesting to study in order to get inspired in the possible symbolic translation (SSTS).

7.3.2 MicroErgo - Concept for Personal Assessment of Occupational Risk in Desk/Office Jobs

A lot of focus has been given to occupational health scenarios during this thesis. Especially for the main projects in which the group was involved. One of these projects is Prevention of Occupational Disorders in the Public Administration with AI (PrevOccupAI), which has the purpose of preventing occupational disorders in office jobs, namely from the public administration. One of the ideas conceptualized during the project was a self-assessment tool for office workers, based on the idea of *microCovid* [42]. The purpose was to help create more awareness about the biomechanical, environmental and mental occupational variables that affect our health. This would be a beneficial approach for any company to self-assess their occupations or even for remote workers who are not always aware if their desk setup is good or not for their biomechanical health, for example. The work can be found here [161].

7.3.3 In using Direct Measures for Occupational Health Assessment

The methods studied and developed in this thesis are general and applicable to any type of time series. This means that these are applicable to direct measures from the occupational domain for information retrieval, as showed on the last section of the previous chapter. We showed that this context was always considered and highly influenced by problems from industry and office/desk jobs.

During this period, a complete understanding of occupational variables was made. This was essential to understand the sensors that could be interesting to use to monitor these variables. Only inertial variables were used to perform a motion capture of upper body segments, which would give most of the angular information needed to study postural variables present on EAWS (this was how Dataset 14, *VAOD*, Section A.1.13 was acquired and more details can be found there.). The usage of direct measures in this context helped in understanding the level of risk a specific workstation represents for a worker. For instance, it is possible to understand for a specific working cycle, which percentage of time it has a high, medium and low risk (using standard ergonomic measures

from RULA). It was also possible to conclude that the same workstation is performed differently by workers with different anthropometric features, which means that a specific workstation should not be considered to have the same risk score for all workers. These conclusions can be found at [167].

The usage of direct measures in this context is therefore highly valuable, considering that standard methods can be used to (mostly) automatically calculate risk scores for each worker and each workstation. Using these measures, a specific workstation can be studied in detail, by means of understanding which processes contribute with the highest risk, for example. Another scenario involves studying how to adapt new workstations to improve productivity or reduce occupational risk. In either case, these direct measures can be used to measure the risk of the processes that were added/removed/modified to the workstation being proposed and understand if it truly is beneficial or not.

The value of direct measures is also related to the existing and continuously increasing knowledge in data mining. Methods, such as the ones developed in this work, can be used to extract relevant information from this data. As we showed, segmentation and pattern search are examples of possible mechanisms for information retrieval in these datasets. Specific *known* shapes can be searched with query-based mechanisms, either by text or *subsequences* used as examples. At some point, supervised learning methods can be made to create working profiles for workstations and workers that consider differences in anthropometric features, specific types of processes and the associations between these.

Finally, another possible usage of these direct measures, would be to design automatically job rotation schedules that use this personal and individual information. As we showed previously, an algorithm was developed with this purpose, but the measures considered were from EAWS standards, which, as reflected in [167], do not consider differences between workers. This provides an additional level of detail that could help to better assign workers to workstations based on their level of capacity.

7.3.4 Support in Motion Training

In industrial tasks, beginner workers have to learn many tasks in multiple workstations. The learning ability differs from worker to worker depending on multiple factors. Several of these tasks have a *standard* motion, with a specific pattern. This *standard* typically follows the optimization for productivity and occupational health. Having a tool that could be used to understand which are the areas of the signal where the motions differ the most and give a real time feedback was an idea that was explored during my stay at the Universität Bremen, being able to develop a tool that asks for two different subjects to perform the same motion and compares it. The videos of both executions is recorded and aligned following the DTW distance. Some highlights of this work can be found at `https://github.com/JmdRodrigues/PhDThesis/tree/main/ProjectCode/LabLinking`.

7.3.5 Volatile Organic Compounds Classification

A Master Thesis in collaboration with the Biomolecular Engineering Group, from the Chemistry Department of the NOVA University of Lisbon. We developed the first version of HeaRTS for the classification of Volatile Organic Compounds (VOC)s. This resulted in a publication that can be found here [7].

7.4 List of Publications

The knowledge gathered during my stay at the Biosignals group from LIBPhys-UNL has been been disseminated *via* scientific publications. In addition, several research collaborations were made that also resulted in collaborative publications. The outcomes are hereby presented.

7.4.1 Journal Publications

- Rodrigues, João, Hui Liu, Duarte Folgado, David Belo, Tanja Schultz, and Hugo Gamboa. 2022. "Feature-Based Information Retrieval of Multimodal Biosignals with a Self-Similarity Matrix: Focus on Automatic Segmentation" Biosensors 12, no. 12: 1182. https://doi.org/10.3390/bios12121182

- Mollaei, Nafiseh, Carlos Fujao, Joao Rodrigues, Catia Cepeda, and Hugo Gamboa. 'Occupational Health Knowledge Discovery Based on Association Rules Applied to Workers' Body Parts Protection: A Case Study in the Automotive Industry'. Computer Methods in Biomechanics and Biomedical Engineering 0, no. 0 (2022): 1–14. https://doi.org/10.1080/10255842.2022.2152678.

- Nunes, Maria Lua, Duarte Folgado, Carlos Fujão, Luís Silva, João Rodrigues, Pedro Matias, Marília Barandas, André V. Carreiro, Sara Madeira, and Hugo Gamboa. 'Posture Risk Assessment in an Automotive Assembly Line Using Inertial Sensors'. IEEE Access 10 (2022): 83221–35. https://doi.org/10.1109/ACCESS.2022.3196473.

- Mollaei, Nafiseh, Carlos Fujao, Luis Silva, Joao Rodrigues, Catia Cepeda, and Hugo Gamboa. 2022. "Human-Centered Explainable Artificial Intelligence: Automotive Occupational Health Protection Profiles in Prevention Musculoskeletal Symptoms" International Journal of Environmental Research and Public Health 19, no. 15: 9552. https://doi.org/10.3390/ijerph19159552

- Gamboa, Patricia, Rui Varandas, João Rodrigues, Cátia Cepeda, Cláudia Quaresma, and Hugo Gamboa. 2022. "Attention Classification Based on Biosignals during Standard Cognitive Tasks for Occupational Domains" Computers 11, no. 4: 49. https://doi.org/10.3390/computers11040049.

- Assunção, Ana, Nafiseh Mollaei, João Rodrigues, Carlos Fujão, Daniel Osório, António P. Veloso, Hugo Gamboa, and Filomena Carnide. 'A Genetic Algorithm Approach to Design Job Rotation Schedules Ensuring Homogeneity and Diversity of Exposure in the Automotive Industry'. Heliyon 8, no. 5 (2022): e09396. https://doi.org/10.1016/j.heliyon.2022.e09396.

- Ramos, G., J. R. Vaz, G. V. Mendonça, P. Pezarat-Correia, J. Rodrigues, M. Alfaras, H. Gamboa, "Fatigue Evaluation through Machine Learning and a Global Fatigue Descriptor", Journal of Healthcare Engineering, vol. 2020, Article ID 6484129, 18 pages, 2020. https://doi.org/10.1155/2020/6484129.

- Rodrigues, João, Duarte Folgado, David Belo, and Hugo Gamboa. 'SSTS: A Syntactic Tool for Pattern Search on Time Series'. Information Processing and Management 56, no. 1 (2019): 61–76. https://doi.org/10.1016/j.ipm.2018.09.001.

- Belo, David, João Rodrigues, João R. Vaz, Pedro Pezarat-Correia, and Hugo Gamboa. "Biosignals Learning and Synthesis Using Deep Neural Networks - Biomedical Engineering Online." BioMed Central. BioMed Central, September 25, 2017. https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-017-0405-0#citeas.

- Rodrigues, João, David Belo, and Hugo Gamboa. 'Noise Detection on ECG Based on Agglomerative Clustering of Morphological Features'. Computers in Biology and Medicine 87 (2017): 322–34. https://doi.org/10.1016/j.compbiomed.2017.06.009.

7.4.2 Book Chapters

- Santos, Sara; Folgado, Duarte; Rodrigues, João; Mollaei, Nafiseh; Fujão, Carlos; Gamboa, Hugo. "Exploring Inertial Sensor Fusion Methods for Direct Ergonomic Assessments". In Communications in Computer and Information Science, 289-303. Springer International Publishing, 2021;

- Gamboa, Patricia; Quaresma, Cláudia; Varandas, Rui; Canhão, Helena; de Sousa, Rute Dinis; Rodrigues, Ana; Jacinto, Sofia; et al. "Design of an Attention Tool Using HCI and Work-Related Variables". In IFIP Advances in Information and Communication Technology, 262-269. Portugal: Springer International Publishing, 2021;

- Rodrigues, João; Gamboa, Hugo; Mollaei, Nafiseh; Osório, Daniel; Assunção, Ana; Fujão, Carlos; Carnide, Filomena. "A Genetic Algorithm to Design Job Rotation Schedules with Low Risk Exposure". In IFIP Advances in Information and Communication Technology, 395-402. Portugal: Springer International Publishing, 2020.

- Cepeda, Catia; Rodrigues, Joao; Dias, Maria Camila; Oliveira, Diogo; Rindlisbacher, Dina; Cheetham, Marcus; Gamboa, Hugo. "Mouse Tracking Measures and Movement Patterns with Application for Online Surveys". In Machine Learning and Knowledge Extraction, 28-42. Springer International Publishing, 2018.

7.4.3 Conference Proceedings

- Shahcheraghi, M.; Mercer, R.; Rodrigues, J.; Der, A.; Gamboa, H.; Zimmerman, Z.; and Keogh, E.; "Scaling Time Series Similarity Matrices to Massive Data", IEEE International Conference on Data Mining (ICDM), Orlando, FL, USA, December 2022.

- Silva, S.; Cepeda, C.; Rodrigues, J.; Probst, P. and Gamboa, H. (2022). Assessing Occupational Health with a Cross-platform Application based on Self-reports and Biosignals. In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF, ISBN 978-989-758-552-4; ISSN 2184-4305, pages 549-556. DOI: 10.5220/0010846700003123

- Alves, R.; Rodrigues, J.; Ramou, E.; Palma, S.; Roque, A. and Gamboa, H. (2022). Classification of Volatile Compounds with Morphological Analysis of e-nose Response. In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS, ISBN 978-989-758-552-4; ISSN 2184-4305, pages 31-39. DOI: 10.5220/0010827200003123

- Mollaei, N.; Cepeda, C.; Rodrigues, J. and Gamboa, H. (2022). Biomedical Text Mining: Applicability of Machine Learning-based Natural Language Processing in Medical Database. In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS, ISBN 978-989-758-552-4; ISSN 2184-4305, pages 159-166. DOI: 10.5220/0010819500003123

- J. Rodrigues et al., "microErgo: A Concept for an Ergonomic Self-Assessment Tool," 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII), 2021, pp. 1-6, doi: 10.1109/ICBSII51839.2021.9445156

- J. Rodrigues, P. Probst and H. Gamboa, "TSSummarize: A Visual Strategy to Summarize Biosignals," 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII), 2021, pp. 1-6, doi: 10.1109/ICBSII51839.2021.9445154.

- Santos, A.; Rodrigues, J.; Folgado, D.; Santos, S.; Fujão, C. and Gamboa, H. (2021). Self-Similarity Matrix of Morphological Features for Motion Data Analysis in Manufacturing Scenarios. In Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS, ISBN 978-989-758-490-9; ISSN 2184-4305, pages 80-90. DOI: 10.5220/0010252800800090

- Rodrigues, J.; Gamboa, H.; Kublanov, V.; Dolganov, A.. "Storage of Biomedical Signals: Comparative Review of Formats and Databases". Paper presented in Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Yekaterinburg, 2019.

7.4.4 Planned Publications

- Novelty Function-Based Automatic Activity Segmentation of Multimodal Biosignals Acquired from Wearable Sensors (in partnership with the Cognitive Systems Lab - University of Bremen)
- HeaRTS - Human Readable Time Series Classification
- Towards Natural Language Search and Exploration of Automotive Telemetry (in partnership with University of California Riverside)
- LabLinking for Remote Motion Synchrony and Teaching (in partnership with the Cognitive Systems Lab - University of Bremen)

7.4.5 Methods

In this thesis, we contributed to the state of the art with several methods, hereby listed.

- TSSummarize: Summarization of Time Series - Using the SSM to provide relevant feedback on how the time series are structured and segments are related, towards automatic and unsupervised annotation of time series.
- SSTS: Synthatic Search on Time Series - performing search on a symbolic representation of times series with regular expressions.
- HeaRTS: Human Readable Time Series - Higher level classification process of time series with visual and possible keyword feedback on data differences.
- QuoTS: *Where* on Time Series? - Text-based query search on word-feature vectors.
- Unsupervised Automatic Annotation: Using the SSM to search for segmentation points on any time series, including novelty segmentation and periodic segmentation, and automatically cluster the segments into similarity groups.

7.4.6 Projects

- Project Operator (2020-2021): I was able to participate on the design of the proposal and contribute in several tasks. Most of my contributions involved (1) the research and investigation of sensing devices that could be beneficial for the monitoring of risk factors variables during work, (2) designing acquisition protocols with specific

research questions, which involved to define the acquisition setup and which activities to perform, (3) participate in the acquisition sessions, and (4) suggesting possible paths for the analysis of the data.

- Project PrevoccupAI (2020-2021): I participated on the design of this proposal and gave contributions at the initial stages of the project. My involvement regarded (1) the research of sensing equipment to be used in desk occupational scenarios, (2) design acquisition protocols and (3) give advice in how to process motion and muscle signals.

7.4.7 Peer Reviewed Works

The content of this thesis used peer-reviewed sections of the listed publications above. The sections that used part of this text are listed as follows:

Chapter	Publication
2	Essential definitions use the text from the peer-reviewed publication [157]
3	The literature review text used on several peer-reviewed works published by the author was used in several of these sections, namely [157, 162, 7, 168, 159]
4	The content was based on the peer-reviewed publication [157]
5	Part of the content and text was based on the peer-reviewed publication [162]
6	Text and results are based on the peer-reviewed publications [157, 168, 162, 7]

7.4.8 Awards

7.4.8.1 Fullbright

I was awarded in 2021 a Fullbright scholarship to pursue a research project at the Computer Science Department of the University of California, Riverside (UCR). The exchange program was made under the supervision of Prof. Eamonn Keogh. It made possible a close collaboration in the development of QuoTS. We are now working in the submission of a conference paper and a journal paper on this topic.

7.4.8.2 Best Paper Award

The best paper award for the category was awarded to the publication "MicroErgo: A Concept for Self-Assessment of Occupational Risk" at the Seventh International Conference on Biosignals, Images and Instrumentation conference, 2021.

7.4.8.3 Hanse-Wissenschaftskolleg Scholarship for Advanced Studies

I was awarded in 2022 a scholarship at the Hanse-Wissenschaftskolleg: Center for Advanced Studies. During this period, I was able to collaborate with the Cognitive Systems Lab from the Universität Bremen, supervised by Prof. Tanja Schultz.

7.5 Future Work

7.5.1 Overall Improvements to compute the SSM and Segmentation Process

The SSM was computed by using the standard approach found at [133]. The results show promise in using this strategy for several tasks. We believe there are improvements that can be made in the feature extraction process, and contribute to a better SSM. Currently, neither a dimensionality reduction method, such as Principal Component Analysis (PCA), nor a feature selection process, are performed prior to computing the SSM. Performing these might be a way of improving the general matrix representation and remove features that do not contribute to explain the similarity between subsequences. Additionally, feature stacking has been proved to help in increasing the accuracy of feature-based classification of time series [66]. This means that it should better represent differences between subsequences and should be tested on the current approach.

Another relevant improvement would be to find a way to minimize the number of parameters used for the novelty segmentation task. Currently, the size of the sliding window that extracts features and the size of the sliding kernel that computes the novelty function are independent. It would be interesting to reduce the number of variables by finding a relationship between the kernel and moving window, make the sizes adaptive based on a specific metric or perform a training process where these parameters are computed for a specific domain and kept the same for future iterations (as we showed the values of these parameters are close together for the same dataset and type of task).

The general idea when this method was used, was that changes on the signal should be represented by a change in the overall set of features. This leads to the idea that different changes can be associated with a specific group of features. In the future, we should study which are the features that better describe a specific type of change.

7.5.2 Benchmarking Multi-Segmentation Tasks in Multimodal Time Series

While working in the topic of segmentation, we did our best to make a fair validation. It was very difficult to find data with multimodal information and that could have multiple events. We initially found one benchmark (*ATCPD*) with real-world time series and multiple events, and more recently found another benchmark of univariate time series, with mostly single events (*UCRS*). We find that more datasets should be available, because in the *ATCPD* Dataset, changes are mainly focused in changes of statistical properties, while in the *UCRS* Dataset, regime changes are covered, but in most of the signals, only

one event is searched, which makes the task easier than what is normally found in data acquired from longer sessions, as in the signals we used in this work. We also think that a better standardization of the validation process should be done. Effectively, the validation we used with TP, FP and FN is typically employed, but might be unfair in cases where an event is not considered because it is at the border of the tolerance zone. However, when using only distance measures, we might be unfairly exaggerating the penalization for cases where FP are, and if the data is not perfectly labelled, wrongly evaluating the used methods.

We believe this topic could be more deeply studied and standardized for the community. This includes finding a benchmark with different types of tasks, different types of changes in the signal, different types of data, include uni- and multi- modal cases, as well as multi- and uni-variate cases. In addition, it should cover tasks with different difficulties in terms of how many events have to be found, but also in terms of difficult it is to find the event. This could be a very good contribution to the community.

7.5.3 Unsupervised Automatic Segmentation and Labelling of Time Series

We have already mentioned in a previous chapter that the proposed methods help moving towards unsupervised and automatic segmentation and labelling of time series. We have showed that the segmentation process is well performed, but lacked the validation for automatic labelling. We believe that using the similarity profiles after a first segmentation process would provide this ability of automatically returning a completely segmented and annotated signal. In the future, we will test this approach on public datasets.

7.5.4 Hierarchical Segmentation of Time Series

When analyzing a long time series visually, the user has to zoom-in and zoom-out to search for areas of interest. We believe that using the proposed segmentation method we could perform a hierarchical segmentation, that is, apply multiple segmentation stages, with different window lengths. This would provide a multi-layered set of information that can highlight the areas of interest in different *zoom* levels.

As a pilot for this method, the user could define the number of hierarchies and corresponding window sizes to perform this process, but ideally, the process would be performed with an adaptive sliding window, that alters the dimensions based on a specific metric. This process would be helpful for long time series, with highly variable information along time.

7.5.5 Query-based Segmentation

Having this methodology to perform segmentation of time series, opens many research paths. One that we believe could be relevant happens in cases where we could write a query that would represent the segmentation pattern we intend to find. A typical

task involves recording data in a methodological, systematic and recurring way, asking multiple participants to perform the same tasks in the same sequence. Knowing the structure in which the data was recorded could be used as hints for the segmentation process. Let us imagine a typical sequence of activities being recorded, such as `walking - running - jumping`, and this was repeated three times. We could write a query that could be used to segment such sequence: `(W R J){3}`. We could use this query as the *hint* given by the user, since it tells the method that it should search for three different regimes (`W`, `R` and `J`) and that it is repeated three times. We believe this could be a relevant contribution to the state of the art in terms of segmentation search.

7.5.6 Periodic Segmentation

The current approach for periodic segmentation is not able to search for the off-diagonals (paths) that are used for this purpose. We believe a better approach could be performed if the paths were highlighted. For this purpose we can perform image processing filters and then extract the resulting paths. Identifying their beginning would be the best way to find the initial sample of a period.

7.5.7 Online Unsupervised Segmentation

This work has not discussed the application of the proposed method for online purposes. We believe the method can be adapted for both novelty segmentation and periodic segmentation. Of course that if the entire SSM is computed over time with continuous incoming data, the memory required for this process would not be enough and the algorithm would fail very quickly. For this purpose, the method should be adapted by only keeping samples from the segmentation point forward (with a fixed buffer size), and compare the next incoming samples with the kept ones until a relevant change is identified or a new off-diagonal starts. After this, the previous samples kept on a buffer up to the segmentation point can be erased and the method can search for the next relevant change point. In the future, an online version should be developed.

7.5.8 Tool for Time Series Profiling

Currently, we introduced *TSSummarize*, which provides a summarization of the time series based on segmentation points and similarity profiles. We believe the development of an interactive tool that has an internal report on the time series, such as, statistical patterns on the segments, number of segments, how similar they are, percentage of time each segment is represented on the signal, level of periodicity, how many periods are there in each segment, presence of anomalies/discords or motifs, among other measures. In the future, this should be considered.

7.5.9 SSTS Improvements and Further Applications

The current SSTS method has several improvements to be made. Some of them have been introduced when developing HeaRTS, but others still require additional research. The fact that we are performing a symbolic representation gives the opportunity to use compression techniques typically used for text, such as the run length encoding (RLE). Having a compressed representation can make the search process faster. In combination with this, it would be interesting to include the higher level translation performed in HeaRTS, which has standard queries for standard structures (peaks, up, plateau, etc...). Using a compression of the time series can be beneficial for HeaRTS to speed up the text representation process and the hyperparameter tuning. A regex query is a text pattern, which is convenient to express a general pattern. However, it is sometimes difficult to generalize. The fact that the regex is not flexible makes this very brittle to patterns that we are looking for but have a slight difference with our text pattern. For now, this flexibility can be introduced with pre-processing/simplification of the data. In the future, a flexible search process, based on a meta-regex mechanism should be developed to perform a less brittle search.

SSTS has a fertile ground where other domains can benefit from these ideas. We have showed applications regarding pattern search and classification, but we believe there are other approaches to explore with SSTS, namely for time series generation, editing, segmentation, among others. The fact that we are searching for patterns makes it convenient to perform interactive adaptations to the data. Several methods have been thought for the *edition* of time series subsequences found with the text pattern, for instance *ssts.annotate*, *ssts.split*, *ssts.modify*, *ssts.replace*, *ssts.reverse*, *ssts.repeat*, *ssts.recursive*. Some of these functions are inspired in text edition mechanisms, others are used for general edition processes. We find that having a tool that could be used to edit, search or adapt a signal could have its benefits, namely to generate datasets for *toy* problems. There are even public libraries to generate pattern sequences based on regex (e.g. *Xeger* [142]). If using a regex pattern based on a specific signal, we could generate a multitude of different signals with small variations.

We have showed that SSTS can be used in combination with existing NLP methods for classification processes and pattern search. We believe that with the current rise in NLP knowledge, a symbolic representation of time series can be useful to design novel ways of extracting information from time series. For instance, several methods are available for text topic modeling, such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF). These methods could be directly tested with the symbolic/textual representation of the time series. Other strategies, that rely in neural networks, such as *bert* and *transformers* could be explored for time series problems, namely for time series classification and generation. We believe these approaches can even be more relevant regarding interpretability and explainability. These are complex problems with time series, and having text as a medium of communication between analysts and

the time series could improve current approaches on this domain. This was explored with HeaRTS, but the process was only evaluating the differences between the data and not explaining why the classifier selected a specific class.

Regarding the topic of classification with HeaRTS, we believe it should be adapted to work with multidimensional data. This could be made by combining several classifiers for each dimension and then combine the classification results by a standard voting method. In addition, HeaRTS could be used in combination with other classifiers as it explores other properties that distances such as ED and DTW don't.

7.5.10 Further Developments for QuoTS

We have introduced a novel method for pattern search on time series with word-feature vectors. This approach is more expressive and provides an easy way to search for patterns. In the future, this expressiveness should be measured with a study group. A group should develop a solution for a problem without using *quots*, and then use QuoTS to solve the same problem. The average time in solving the problem would be measured to associate with the expressiveness. In the long run, this method should be considered for non-experts in time series as well, to understand what could be improved to make the process more expressive for non-experienced users.

In terms of the method itself, more keywords and operators could be introduced. Some of the keywords could even be represented by several features that contribute for the used keyword. The fact that a static window is used, should be improved. For example, SSTS can find patterns with any size, while QuoTS searches for patterns on a fixed window size. Another improvement is related with the feedback given. QuoTS could have a set of methods that help get intuition over what the keywords mean in the time series. For instance, the user could highlight a segment of the time series and the keywords with higher value would be presented.

Bibliography

[1] A. Q. Abbasi and W. A. Loun. "Symbolic time series analysis of temporal gait dynamics". In: *Journal of Signal Processing Systems* 74.3 (2014), pp. 417–422. ISSN: 19398115. DOI: 10.1007/s11265-013-0836-1 (cit. on p. 37).

[2] R. P. Adams and D. J. C. Mackay. "Bayesian Online Changepoint Detection". In: *arXiv: Machine Learning* (2007) (cit. on p. 33).

[3] R. Agrawal, C. Faloutsos, and A. Swami. "Efficient similarity search in sequence databases". In: *Foundations of Data Organization and Algorithms*. Ed. by D. B. Lomet. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 69–84. ISBN: 978-3-540-48047-1 (cit. on p. 11).

[4] R. Agrawal et al. "Querying Shapes of Histories". In: *Proceedings of the 21th International Conference on Very Large Data Bases*. VLDB '95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 502–514. ISBN: 1558603794 (cit. on pp. 20, 38).

[5] K. R. Agres et al. *Music, Computing, and Health: A roadmap for the current and future roles of music technology for health care and well-being*. 2021-02. DOI: 10.1177/2059204321997709. URL: osf.io/mgjwv (cit. on p. 42).

[6] M. Aickin and H. Gensler. "Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods." In: *American Journal of Public Health* 86.5 (1996). PMID: 8629727, pp. 726–728. DOI: 10.2105/AJPH.86.5.726. eprint: https://doi.org/10.2105/AJPH.86.5.726. URL: https://doi.org/10.2105/AJPH.86.5.726 (cit. on p. 30).

[7] R. Alves. et al. "Classification of Volatile Compounds with Morphological Analysis of e-nose Response". In: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS*, INSTICC. SciTePress, 2022, pp. 31–39. ISBN: 978-989-758-552-4. DOI: 10.5220/0010827200003123 (cit. on pp. 131, 135).

[8] S. Aminikhanghahi and D. J. Cook. "A Survey of Methods for Time Series Change Point Detection". In: *Knowledge and Information Systems* 51 (2017-05). ISSN: 0219-3116. DOI: 10.1007/s10115-016-0987-z. URL: https://doi.org/10.1007/s10115-016-0987-z (cit. on pp. 1, 32, 33).

[9] D. Anguita et al. "A Public Domain Dataset for Human Activity Recognition using Smartphones". In: 2013-01 (cit. on p. 165).

[10] D. Anguita et al. "Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine". In: *Ambient Assisted Living and Home Care*. Ed. by J. Bravo, R. Hervás, and M. Rodríguez. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 216–223. ISBN: 978-3-642-35395-6 (cit. on p. 165).

[11] A. Antoniou and C. Vorlow. "Recurrence plots and financial time series analysis". In: 10 (2000-01), pp. 131–145 (cit. on p. 34).

[12] A. Assunção et al. "A genetic algorithm approach to design job rotation schedules ensuring homogeneity and diversity of exposure in the automotive industry". In: *Heliyon* 8 (5 2022-05). DOI: 10.1016/j.heliyon.2022.e09396. URL: https://doi.org/10.1016/j.heliyon.2022.e09396 (cit. on pp. 4, 129).

[13] I. Auger. "Algorithms for the optimal identification of segment neighborhoods". In: *Bltn Mathcal Biology* 51 (1989), pp. 39–54. DOI: https://doi.org/10.1007/BF02458835 (cit. on p. 33).

[14] A. Bagnall and E. Keog. *Time Series Classification Website*. 2022. URL: http://www.timeseriesclassification.com/urldate%20=%20%7B4-06-2022%7D (cit. on p. 164).

[15] J. Bai. "Estimating Multiple Breaks One at a Time". In: *Econometric Theory* 13.3 (1997), pp. 315–352. DOI: 10.1017/S0266466600005831 (cit. on p. 33).

[16] M. Barandas et al. "TSFEL: Time Series Feature Extraction Library". en. In: *SoftwareX* 11 (2020-01), p. 100456. ISSN: 23527110. DOI: 10.1016/j.softx.2020.100456. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352711020300017 (visited on 2020-04-03) (cit. on p. 44).

[17] G. Batista et al. "CID: An efficient complexity-invariant distance for time series". In: *Data Mining and Knowledge Discovery* 28 (2013-04). DOI: 10.1007/s10618-013-0312-3 (cit. on pp. 15, 16).

[18] K. Bauters et al. "An automated work cycle classification and disturbance detection tool for assembly line work stations". In: *ICINCO 2014 - Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics*. Vol. 2. SciTePress, 2014, pp. 685–691. ISBN: 9789897580406. DOI: 10.5220/0005024406850691 (cit. on p. 34).

[19] V. Behravan et al. "Rate-adaptive compressed-sensing and sparsity variance of biomedical signals". In: *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 2015, pp. 1–6. DOI: 10.1109/BSN.2015.7299419 (cit. on pp. 1, 168, 169).

[20] J. P. Bello et al. "Content-based Methods for Knowledge Discovery in Music". In: *Springer Handbook on Systematic Musicology*. Ed. by R. Bader. Springer, Berlin, Heidelberg, 2018, pp. 823–840. ISBN: 978-3-662-55002-1. DOI: https://doi.org/10.1007/978-3-662-55004-5_39 (cit. on pp. 36, 42, 46).

[21] D. Belo. "Learning Biosignals using Deep Learning". PhD dissertation. Nova University of Lisbon, 2020 (cit. on pp. 1, 40).

[22] G. M. Bhandari, R. S. Kawitkar, and M. P. Borawake. "Audio Segmentation for Speech Recognition Using Segment Features". In: *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II*. Ed. by S. C. Satapathy et al. Cham: Springer International Publishing, 2014, pp. 209–217. ISBN: 978-3-319-03095-1 (cit. on p. 42).

[23] A. P. L. Bó, M. Hayashibe, and P. Poignet. "Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect". In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2011, pp. 3479–3483. DOI: 10.1109/IEMBS.2011.6090940 (cit. on p. 1).

[24] P. Bonato. *Advances in wearable technology and applications in physical medicine and rehabilitation*. 2005 (cit. on p. 40).

[25] M. Bosc et al. "Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution". In: *NeuroImage* 20.2 (2003), pp. 643–656. ISSN: 1053-8119. DOI: https://doi.org/10.1016/S1053-8119(03)00406-3. URL: http://www.sciencedirect.com/science/article/pii/S1053811903004063 (cit. on p. 1).

[26] G. J. J. van den Burg and C. K. I. Williams. "An Evaluation of Change Point Detection Algorithms". In: (2020-05). DOI: arXiv:2003.06222 (cit. on pp. 9, 32, 33, 85, 86, 88, 89, 169, 170, 185).

[27] F. Caputo et al. "IMU-Based Motion Capture Wearable System for Ergonomic Assessment in Industrial Environment". In: *Springer Nature* (2019), pp. 215–225. DOI: 10.1007/978-3-319-94619-1_21 (cit. on p. 4).

[28] P. Catalfamo, S. Ghoussayni, and D. Ewins. "Gait Event Detection on Level Ground and Incline Walking Using a Rate Gyroscope". In: *Sensors* 10.6 (2010), pp. 5683–5702. ISSN: 1424-8220. DOI: 10.3390/s100605683. URL: https://www.mdpi.com/1424-8220/10/6/5683 (cit. on p. 33).

[29] J. Chen et al. "Wearable sensors for reliable fall detection". In: *2005 IEEE engineering in medicine and biology 27th annual conference*. IEEE. 2006, pp. 3551–3554 (cit. on p. 40).

[30] K.-H. Chen et al. "Wearable sensor-based rehabilitation exercise assessment for knee osteoarthritis". In: *Sensors* 15.2 (2015), pp. 4193–4211 (cit. on p. 40).

[31] L. Chen, C. Nugent, and G. Okeyo. "An ontology-based hybrid approach to activity modeling for smart homes". In: *IEEE Transactions on human-machine systems* 44.1 (2013), pp. 92–105 (cit. on p. 40).

[32] L. Chen et al. "Sensor-based activity recognition". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.6 (2012), pp. 790–808 (cit. on p. 40).

[33] M. Chen, S. Mao, and Y. Liu. "Big data: A survey". In: *Mobile networks and applications* 19.2 (2014), pp. 171–209 (cit. on p. 40).

[34] H. Cho and P. Fryzlewicz. "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.2 (2014-07), pp. 475–507. ISSN: 1369-7412. DOI: 10.1111/rssb.12079. URL: http://dx.doi.org/10.1111/rssb.12079 (cit. on p. 1).

[35] K. Chuttani et al. "Diagnosis of cardiac tamponade after cardiac surgery: Relative value of clinical, echocardiographic, and hemodynamic signs". In: *American Heart Journal* 127.4, Part 1 (1994), pp. 913–918. ISSN: 0002-8703. DOI: https://doi.org/10.1016/0002-8703(94)90561-4. URL: https://www.sciencedirect.com/science/article/pii/0002870394905614 (cit. on p. 173).

[36] D. Cook and N. Krishnan. *Activity learning: Discovering, recognizing, and predicting human behavior from sensor data*. 2015-02, pp. 1–257. ISBN: 9781118893760. DOI: 10.1002/9781119010258 (cit. on p. 1).

[37] C. Cortes and V. Vapnik. "Support-Vector Networks". In: *Mach. Learn.* 20.3 (1995-09), pp. 273–297. ISSN: 0885-6125. DOI: 10.1023/A:1022627411411. URL: https://doi.org/10.1023/A:1022627411411 (cit. on p. 24).

[38] D. P. Coutinho, A. L. N. Fred, and M. A. T. Figueiredo. "One-Lead ECG-based Personal Identification Using Ziv-Merhav Cross Parsing". In: 20^{th} *International Conference on Pattern Recognition*. 2010-08, pp. 3858–3861. DOI: 10.1109/ICPR.2010.940 (cit. on p. 37).

[39] D. Crystal. *Making Sense of Grammar*. Pearson Education, 2004-01. ISBN: 9780582848634 (cit. on p. 3).

[40] Z. Cui et al. "Enhancing Interactions for In-Car Voice User Interface with Gestural Input on the Steering Wheel". In: *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI '21. Leeds, United Kingdom: Association for Computing Machinery, 2021, pp. 59–68. ISBN: 9781450380638. DOI: 10.1145/3409118.3475126. URL: https://doi.org/10.1145/3409118.3475126 (cit. on p. 115).

[41] P. Cunningham and S. J. Delany. "K-Nearest Neighbour Classifiers - A Tutorial". In: *ACM Comput. Surv.* 54.6 (2021-07). ISSN: 0360-0300. DOI: 10.1145/3459665. URL: https://doi.org/10.1145/3459665 (cit. on p. 21).

[42] D. Dagradi et al. *microCovid Project*. 2020. URL: https://www.microcovid.org/ (visited on 2022-07-07) (cit. on p. 129).

[43] R. B. Dannenberg and M. Goto. "Music Structure Analysis from Acoustic Signals". In: *Handbook of Signal Processing in Acoustics*. Ed. by D. Havelock, S. Kuwano, and M. Vorländer. New York, NY: Springer New York, 2008, pp. 305–331. ISBN: 978-0-387-30441-0. DOI: 10.1007/978-0-387-30441-0_21. URL: https://doi.org/10.1007/978-0-387-30441-0_21 (cit. on p. 48).

[44] H. A. Dau et al. *The UCR Time Series Archive*. 2018. DOI: 10.48550/ARXIV.1810.07758. URL: https://arxiv.org/abs/1810.07758 (cit. on pp. 109, 110, 164).

[45] A. Ejupi et al. "A Kinect and Inertial Sensor-Based System for the Self-Assessment of Fall Risk: A Home-Based Study in Older People". In: *Hum.-Comput. Interact.* 31.3–4 (2016-07), pp. 261–293. ISSN: 0737-0024. DOI: 10.1080/07370024.2015.1085309. URL: https://doi.org/10.1080/07370024.2015.1085309 (cit. on p. 1).

[46] H. G. Espinosa, J. Lee, and D. A. James. "THE INERTIAL SENSOR: A BASE PLATFORM FOR WIDER ADOPTION IN SPORTS SCIENCE APPLICATIONS." In: *Journal of Fitness Research* 4.1 (2015) (cit. on pp. 1, 40).

[47] J. Faouzi and H. Janati. "pyts: A Python Package for Time Series Classification". In: *Journal of Machine Learning Research* 21.46 (2020), pp. 1–6. URL: http://jmlr.org/papers/v21/19-763.html (cit. on p. 13).

[48] D. Folgado et al. "TSSEARCH: Time Series Subsequence Search Library". In: *SoftwareX* 18 (2022), p. 101049. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2022.101049 (cit. on p. 33).

[49] J. Foote. "Automatic audio segmentation using a measure of audio novelty". In: *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*. Vol. 1. 2000, 452–455 vol.1. DOI: 10.1109/ICME.2000.869637 (cit. on p. 47).

[50] P. S. Foundation. *10.3. operator - Standard Operators as functions*. 2017. URL: https://docs.python.org/3/library/operator.html (visited on 2017-09-27) (cit. on p. 106).

[51] P. S. Foundation. *regex 2017.09.23*. 2017. URL: https://pypi.python.org/pypi/regex/ (visited on 2017-10-03) (cit. on p. 67).

[52] J. Friedl. *Mastering Regular Expressions*. 3^{rd}. O'Rilley Media, Inc, 2006 (cit. on p. 25).

[53] M. Gadaleta et al. "Deep Learning Techniques for Improving Digital Gait Segmentation". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019-07. DOI: 10.1109/embc.2019.8856685. URL: https://doi.org/10.1109%2Fembc.2019.8856685 (cit. on p. 33).

[54] M. Gadaleta et al. "Deep Learning Techniques for Improving Digital Gait Segmentation". In: (2019-07) (cit. on p. 33).

[55] P. Gamboa et al. "Attention Classification Based on Biosignals during Standard Cognitive Tasks for Occupational Domains". In: *Computers* 11.4 (2022). ISSN: 2073-431X. DOI: 10.3390/computers11040049. URL: https://www.mdpi.com/2073-431X/11/4/49 (cit. on p. 4).

[56] Y. Ganesan, S. Gobee, and V. Durairajah. "Development of an upper limb exoskeleton for rehabilitation with feedback from EMG and IMU sensor". In: *Procedia Computer Science* 76 (2015), pp. 53–59 (cit. on p. 40).

[57] S. Gharghabi et al. "Domain agnostic online semantic segmentation for multi-dimensional time series". In: *Data Mining and Knowledge Discovery* 33 (2019), pp. 96–130. ISSN: 1573-756X. URL: https://doi.org/10.1007/s10618-018-0589-3 (cit. on pp. 34, 55, 172).

[58] S. Gharghabi et al. "Matrix Profile VIII: Domain Agnostic Online Semantic Segmentation at Superhuman Performance Levels". In: *2017 IEEE International Conference on Data Mining (ICDM)*. 2017, pp. 117–126. DOI: 10.1109/ICDM.2017.21 (cit. on p. 18).

[59] S. Gharghabi et al. "Matrix Profile XII: MPdist: A Novel Time Series Distance Measure to Allow Data Mining in More Challenging Scenarios". In: *2018 IEEE International Conference on Data Mining (ICDM)* (2018), pp. 965–970 (cit. on p. 34).

[60] A. L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23 (2000-06). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215, e215–e220 (cit. on pp. 119, 165, 168, 169).

[61] A. L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet". In: *Circulation* 101.23 (2000), e215–e220. ISSN: 0009-7322. DOI: 10.1161/01.CIR.101.23.e215. eprint: http://circ.ahajournals.org/content/101/23/e215.full.pdf (cit. on p. 53).

[62] D. H. Seidel et al. "Quantitative Measures of Physical Risk Factors Associated with Work-Related Musculoskeletal Disorders of the Elbow: A Systematic Review". In: *International Journal of Environmental Research and Public Health* 16 (2019-01), p. 130. DOI: 10.3390/ijerph16010130 (cit. on p. 4).

[63] C. L. Hamblin. "Translation to and from Polish Notation". In: *The Computer Journal* 5.3 (1962), pp. 210–213. DOI: 10.1093/comjnl/5.3.210. eprint: /oup/backfile/content_public/journal/comjnl/5/3/10.1093/comjnl/5.3.210/2/5-3-210.pdf. URL: +%20http://dx.doi.org/10.1093/comjnl/5.3.210 (cit. on p. 63).

[64] S. Hamdi, A. Ben Abdallah, and M. H. Bedoui. "Real time QRS complex detection using DFA and regular grammar". In: *BioMedical Engineering OnLine* 16.1 (2017-02), p. 31. ISSN: 1475-925X. DOI: 10.1186/s12938-017-0322-2. URL: https://doi.org/10.1186/s12938-017-0322-2 (cit. on p. 37).

[65] J. Han, M. Kamber, and J. Pei. "2 - Getting to Know Your Data". In: *Data Mining (Third Edition)*. Ed. by J. Han, M. Kamber, and J. Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 39–82. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00002-2. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000022 (cit. on p. 17).

[66] Y. Hartmann, H. Liu, and T. Schultz. "Feature Space Reduction for Human Activity Recognition based on Multi-channel Biosignals". In: *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*. INSTICC. SCITEPRESS - Science and Technology Publications, 2021, pp. 215–222. ISBN: 978-989-758-490-9. DOI: 10.5220/0010260802150222 (cit. on p. 136).

[67] Y. Hartmann, H. Liu, and T. Schultz. "Interactive and Interpretable Online Human Activity Recognition". In: *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE. 2022, pp. 109–111 (cit. on p. 40).

[68] T. Heldt et al. "Circulatory response to passive and active changes in posture". In: *Computers in Cardiology, 2003*. 2003, pp. 263–266. DOI: 10.1109/CIC.2003.1291141 (cit. on pp. 53, 164, 165).

[69] R. A. Henning, S. L. Sauter, and E. F. Krieg. "Work rhythm and physiological rhythms in repetitive computer work: Effects of synchronization on well-being". In: *International Journal of Human–Computer Interaction* 4.3 (1992), pp. 233–243. DOI: 10.1080/10447319209526040. eprint: https://doi.org/10.1080/1044

7319209526040. URL: https://doi.org/10.1080/10447319209526040 (cit. on p. 4).

[70] A. Holzinger. "Human-Computer Interaction and Knowledge Discovery (HCI-KDD): What Is the Benefit of Bringing Those Two Fields to Work Together?" In: *Availability, Reliability, and Security in Information Systems and HCI*. Ed. by A. Cuzzocrea et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 319–328. ISBN: 978-3-642-40511-2 (cit. on p. 2).

[71] A. Holzinger and G. Pasi. "Introduction to the special issue on "interactive data analysis"". In: *Information Processing and Management* 51.2 (2015), pp. 141–143. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2014.11.002. URL: http://www.sciencedirect.com/science/article/pii/S0306457314001101 (cit. on p. 2).

[72] A. Holzinger and M. Zupan. "KNODWAT: A scientific framework application for testing knowledge discovery methods for the biomedical domain". In: *BMC Bioinformatics* 14.1 (2013-06), p. 191. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-191. URL: https://doi.org/10.1186/1471-2105-14-191 (cit. on p. 2).

[73] S. L. Horowitz. "A Syntactic Algorithm for Peak Detection in Waveforms with Applications to Cardiography". In: *Commun. ACM* 18.5 (1975-05), pp. 281–285. ISSN: 0001-0782. DOI: 10.1145/360762.360810. URL: http://doi.acm.org/10.1145/360762.360810 (cit. on p. 37).

[74] R. Howard. "Wireless sensor devices in sports performance". In: *IEEE Potentials* 35.4 (2016), pp. 40–42 (cit. on pp. 1, 40).

[75] R. M. Howard, R. Conway, and A. J. Harrison. "A survey of sensor devices: use in sports biomechanics". In: *Sports biomechanics* 15.4 (2016), pp. 450–461 (cit. on pp. 1, 40).

[76] B. Hu, E. Rouse, and L. Hargrove. "Benchmark Datasets for Bilateral Lower-Limb Neuromechanical Signals from Wearable Sensors during Unassisted Locomotion in Able-Bodied Individuals". In: *Frontiers in Robotics and AI* 5 (2018). ISSN: 2296-9144. DOI: 10.3389/frobt.2018.00014. URL: https://www.frontiersin.org/articles/10.3389/frobt.2018.00014 (cit. on p. 171).

[77] Y. Hu et al. "OmicCircos: A Simple-to-Use R Package for the Circular Visualization of Multidimensional Omics Data". In: *Cancer Informatics* 13 (2014). PMID: 24526832, CIN.S13495. DOI: 10.4137/CIN.S13495. eprint: https://doi.org/10.4137/CIN.S13495. URL: https://doi.org/10.4137/CIN.S13495 (cit. on p. 36).

[78] L. Hussain et al. "Symbolic time series analysis of electroencephalographic (EEG) epileptic seizure and brain dynamics with eye-open and eye-closed subjects during resting states". In: *Journal of physiological anthropology* 36.1 (2017), p. 21. ISSN: 18806805. DOI: 10.1186/s40101-017-0136-8 (cit. on p. 37).

[79] S. Imani, S. Alaee, and E. Keogh. "Putting the Human in the Time Series Analytics Loop". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco, USA: Association for Computing Machinery, 2019, pp. 635–644. ISBN: 9781450366755. DOI: 10.1145/3308560.3317308. URL: https://doi.org/10.1145/3308560.3317308 (cit. on pp. 1, 38, 39).

[80] S. Imani, S. Alaee, and E. Keogh. "Qute: Query by Text Search for Time Series Data". In: 2021-01, pp. 412–427. ISBN: 978-3-030-63088-1. DOI: 10.1007/978-3-030-63089-8_27 (cit. on pp. 38, 39).

[81] S. Imani and E. J. Keogh. "Matrix Profile XIX: Time Series Semantic Motifs: A New Primitive for Finding Higher-Level Structure in Time Series". In: *2019 IEEE International Conference on Data Mining (ICDM)* (2019), pp. 329–338 (cit. on p. 34).

[82] S. Imani et al. "Introducing time series snippets: a new primitive for summarizing long time series". In: *Data Min. Knowl. Discov.* (2020), pp. 1713–1743 (cit. on p. 35).

[83] E. S. Irastorza, Xabier, and S. Copsey. *OSH in figures: Work-related musculoskeletal disorders in the EU — Facts and figures*. European Agency for Safety and Health at Work, 2010 (cit. on p. 4).

[84] H. Ismail Fawaz et al. "Deep learning for time series classification: a review". In: *Data Mining and Knowledge Discovery* 33.4 (2019), pp. 917–963 (cit. on pp. 30, 89).

[85] I. Jakob et al. "Robotic and sensor technology for upper limb rehabilitation". In: *PM&R* 10 (2018), S189–S197 (cit. on p. 40).

[86] S. L. Jan Goyvaerts. *Regular Expressions Cookbook*. 2nd ed. O'Reilly Media, Inc, 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2012. ISBN: 978-1-449-31943-4 (cit. on p. 106).

[87] Q. Ji et al. "Real-time gait event detection in a real-world environment using a laser-ranging sensor and gyroscope fusion method". In: *Physiological Measurement* 39.12 (2018), p. 125003 (cit. on pp. 1, 40).

[88] K.S.Fu. "Stochastic Automata, Stochastic Languages and Pattern Recognition". In: *Journal of Cybernetics* 1.3 (1971), pp. 31–49. DOI: 10.1080/01969727108548630. eprint: http://dx.doi.org/10.1080/01969727108548630. URL: http://dx.doi.org/10.1080/01969727108548630 (cit. on p. 37).

[89] M. Kayam et al. "Assessing your algorithm: A program complexity metrics for basic algorithmic thinking education". In: *2016 11th International Conference on Computer Science Education (ICCSE)*. 2016-08, pp. 309–313. DOI: 10.1109/ICCSE.2016.7581599 (cit. on p. 28).

[90] E. Keogh. *How to do good research, get it published in SIGKDD and get it cited*. 2009-07. URL: https://www.cs.ucr.edu/~eamonn/Keogh_SIGKDD09_tutorial.pdf (cit. on p. 29).

[91] E. Keogh and R. Chotirat Ann. "Exact Indexing of Dynamic Time Warping". In: *knowledge and Information Systems* (2005-03). DOI: 10.1007/s10115-004-0154-9. URL: https://doi.org/10.1007/s10115-004-0154-9 (cit. on p. 15).

[92] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. "Towards Parameter-Free Data Mining". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. Seattle, WA, USA: Association for Computing Machinery, 2004, pp. 206–215. ISBN: 1581138881. DOI: 10.1145/1014052.1014077. URL: https://doi.org/10.1145/1014052.1014077 (cit. on pp. 38, 39).

[93] E. Keogh et al. "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases". In: 2001-01. URL: https://www.microsoft.com/en-us/research/publication/dimensionality-reduction-for-fast-similarity-search-in-large-time-series-databases/ (cit. on p. 13).

[94] L. Kirichenko et al. "Two Approaches to Machine Learning Classification of Time Series Based on Recurrence Plots". In: *2020 IEEE Third International Conference on Data Stream Mining and Processing (DSMP)*. 2020, pp. 84–89. DOI: 10.1109/DSMP47368.2020.9204021 (cit. on p. 34).

[95] O. Korn and A. Rees. "Affective effects of gamification: using biosignals to measure the effects on working and learning users". In: 2019-06, pp. 1–10. DOI: 10.1145/3316782.3316783 (cit. on p. 1).

[96] N. Kumar et al. "Time-series Bitmaps: a Practical Visualization Tool for Working with Large Time Series Databases". In: 2005-04. DOI: 10.1137/1.9781611972757.55 (cit. on pp. 35, 36, 57).

[97] S. Kurtek et al. "Segmentation, alignment and statistical analysis of biosignals with application to disease classification". In: *Journal of Applied Statistics* 40.6 (2013), pp. 1270–1288. DOI: 10.1080/02664763.2013.785492. eprint: https://doi.org/10.1080/02664763.2013.785492. URL: https://doi.org/10.1080/02664763.2013.785492 (cit. on p. 33).

[98] S.-J. Lee et al. "Exercise and cardiovascular load in workers with high occupational physical activity". In: *Archives of Environmental and Occupational Health* 75.6 (2020). PMID: 31456490, pp. 339–345. DOI: 10.1080/19338244.2019.1657059. eprint: https://doi.org/10.1080/19338244.2019.1657059. URL: https://doi.org/10.1080/19338244.2019.1657059 (cit. on p. 4).

[99] H. Lefebvre, C. Legner, and M. Fadler. "Data democratization: toward a deeper understanding". In: 2021-09 (cit. on p. 2).

[100] Y. Li and A. Ray. "Unsupervised Symbolization of Signal Time Series for Extraction of the Embedded Information". In: *Entropy* 19.4 (2017). ISSN: 1099-4300. DOI: 10.3390/e19040148. URL: http://www.mdpi.com/1099-4300/19/4/148 (cit. on p. 37).

[101] K. Li. and C. Zhou. "Estimation of Gait Parameters based on Motion Sensor Data". In: *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - BIODEVICES*, INSTICC. SciTePress, 2020, pp. 129–135. ISBN: 978-989-758-398-8. DOI: 10.5220/0008963901290135 (cit. on pp. 1, 40).

[102] J. Lin, R. Khade, and Y. Li. "Rotation-Invariant Similarity in Time Series Using Bag-of-Patterns Representation". In: *J. Intell. Inf. Syst.* 39.2 (2012-10), pp. 287–315. ISSN: 0925-9902. DOI: 10.1007/s10844-012-0196-5. URL: https://doi.org/10.1007/s10844-012-0196-5 (cit. on pp. 38, 39).

[103] J. Lin et al. "Experiencing SAX: a novel symbolic representation of time series". In: *Data Mining and Knowledge Discovery* 15.2 (2007-10), pp. 107–144. ISSN: 1573-756X. DOI: 10.1007/s10618-007-0064-z. URL: https://doi.org/10.1007/s10618-007-0064-z (cit. on pp. 13, 14, 35, 37, 39, 127).

[104] H. Liu. "Biosignal processing and activity modeling for multimodal human activity recognition". PhD thesis. University of Bremen, 2021. DOI: 10.26092/elib/1219 (cit. on pp. 40, 44).

[105] H. Liu, Y. Hartmann, and T. Schultz. "A Practical Wearable Sensor-Based Human Activity Recognition Research Pipeline". In: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF*. 2022, pp. 847–856. ISBN: 978-989-758-552-4. DOI: 10.5220/0010937000003123 (cit. on p. 44).

[106] H. Liu, Y. Hartmann, and T. Schultz. "CSL-SHARE: A Multimodal Wearable Sensor-Based Human Activity Dataset". In: *Frontiers in Computer Science* 3 (2021-10), p. 90. DOI: 10.3389/fcomp.2021.759136 (cit. on pp. 170, 171).

[107] H. Liu, Y. Hartmann, and T. Schultz. "Motion Units: Generalized Sequence Modeling of Human Activities for Sensor-Based Activity Recognition". In: *29th European Signal Processing Conference (EUSIPCO 2021)*. IEEE. 2021. DOI: 10.23919/EUSIPCO54536.2021.9616298 (cit. on p. 40).

[108] H. Liu and T. Schultz. "A Wearable Real-time Human Activity Recognition System using Biosensors Integrated into a Knee Bandage". In: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 1: BIODEVICES*. INSTICC. SciTePress, 2019, pp. 47–55. ISBN: 978-989-758-353-7 (cit. on p. 40).

[109] H. Liu and T. Schultz. "How Long Are Various Types of Daily Activities? Statistical Analysis of a Multimodal Wearable Sensor-Based Human Activity Dataset". In: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF*. 2022, pp. 680–688. ISBN: 978-989-758-552-4 (cit. on pp. 1, 40).

[110] J. Liu et al. "uWave: Accelerometer-based personalized gesture recognition and its applications". In: *Pervasive and Mobile Computing* 5.6 (2009). PerCom 2009, pp. 657–675. ISSN: 1574-1192. DOI: https://doi.org/10.1016/j.pmcj.2009.07.007. URL: https://www.sciencedirect.com/science/article/pii/S1574119209000674 (cit. on pp. 115, 117).

[111] S. Lobov et al. "Latent Factors Limiting the Performance of sEMG-Interfaces". In: *Sensors* 18 (2018-04), p. 1122. DOI: 10.3390/s18041122 (cit. on p. 167).

[112] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. ii).

[113] C. M. Lu and N. J. Ferrier. "Automated Analysis of Repetitive Joint Motion". In: *IEEE Transactions on Information Technology in Biomedicine* 7.4 (2003-12), pp. 263–273. ISSN: 10897771. DOI: 10.1109/TITB.2003.821309 (cit. on p. 34).

[114] C. M. Lu and N. J. Ferrier. "Repetitive Motion Analysis: Segmentation and Event Classification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2 (2004-02), pp. 258–263. ISSN: 01628828. DOI: 10.1109/TPAMI.2004.1262196 (cit. on p. 34).

[115] Y. Lu et al. "Matrix Profile XXIV: Scaling Time Series Anomaly Detection to Trillions of Datapoints and Ultra-Fast Arriving Data Streams". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, pp. 1173–1182. ISBN: 9781450393850. DOI: 10.1145/3534678.3539271. URL: https://doi.org/10.1145/3534678.3539271 (cit. on p. 34).

[116] A. Luttmann et al. *Preventing Musculoskeletal Disorders in the Workplace*. World Health Organization, 2003. ISBN: 92 4 159053 X. URL: https://www.who.int/occupational_health/publications/en/oehmsd3.pdf (cit. on p. 4).

[117] R. G. Lyons. *Understanding Digital Signal Processing*. 1st. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1996. ISBN: 0201634678 (cit. on p. 63).

[118] I. P. Machado et al. "Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization". In: *Information Processing and Management* 51.2 (2015), pp. 204–214. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2014.07.008. URL:

http://www.sciencedirect.com/science/article/pii/S0306457314000685 (cit. on pp. 2, 34).

[119] F. Madrid et al. "Matrix Profile XVI: Efficient and Effective Labeling of Massive Time Series Archives". In: *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2019), pp. 463–472 (cit. on p. 34).

[120] A. Malali et al. "Supervised ECG wave segmentation using convolutional LSTM". In: *ICT Express* 6.3 (2020), pp. 166–169. ISSN: 2405-9595. DOI: https://doi.org/10.1016/j.icte.2020.04.004. URL: https://www.sciencedirect.com/science/article/pii/S2405959520300989 (cit. on p. 32).

[121] R. Malladi, G. P. Kalamangalam, and B. Aazhang. "Online Bayesian change point detection algorithms for segmentation of epileptic activity". In: *2013 Asilomar Conference on Signals, Systems and Computers*. 2013, pp. 1833–1837. DOI: 10.1109/ACSSC.2013.6810619 (cit. on p. 1).

[122] Y. Mangukiya, B. Purohit, and K. George. "Electromyography (EMG) sensor controlled assistive orthotic robotic arm for forearm movement". In: *2017 IEEE Sensors Applications Symposium (SAS)*. IEEE. 2017, pp. 1–4 (cit. on p. 40).

[123] M. Mannino and A. Abouzied. "Expressive Time Series Querying with Hand-Drawn Scale-Free Sketches". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–13. ISBN: 9781450356206. DOI: 10.1145/3173574.3173962. URL: https://doi.org/10.1145/3173574.3173962 (cit. on pp. 20, 38).

[124] H. F. Maqbool et al. "A Real-Time Gait Event Detection for Lower Limb Prosthesis Control and Evaluation". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.9 (2017), pp. 1500–1509. DOI: 10.1109/TNSRE.2016.2636367 (cit. on p. 33).

[125] P. Matias et al. "Time Series Segmentation Using Neural Networks with Cross-Domain Transfer Learning". In: *Electronics* 10 (2021-07), p. 1805. DOI: 10.3390/electronics10151805 (cit. on p. 32).

[126] L. McAtamney and E. Nigel Corlett. "RULA: a survey method for the investigation of work-related upper limb disorders". In: *Applied Ergonomics* 24.2 (1993), pp. 91–99. ISSN: 0003-6870. DOI: https://doi.org/10.1016/0003-6870(93)90080-S. URL: http://www.sciencedirect.com/science/article/pii/000368709390080S (cit. on p. 4).

[127] T. McNab, D. A. James, and D. Rowlands. "iPhone sensor platforms: Applications to sports monitoring". In: *Procedia Engineering* 13 (2011), pp. 507–512 (cit. on pp. 1, 40).

[128] J. J. A. Mendes Jr et al. "Sensor fusion and smart sensor in sports and biomedical applications". In: *Sensors* 16.10 (2016), p. 1569 (cit. on pp. 1, 40).

[129] H. B. Menz and D. R. Bonanno. "Objective measurement of adherence to wearing foot orthoses using an embedded temperature sensor". In: *Medical Engineering & Physics* 88 (2021), pp. 19–24 (cit. on p. 40).

[130] V. MN. Dam and B. Fitzgerald. *Pulsus Paradoxus*. 2022-01. URL: https://www.ncbi.nlm.nih.gov/books/NBK482292/ (cit. on pp. 55, 173).

[131] G. B. Moody, W. K. Muldrow, and R. G. Mark. "A Noise Stress for Arrythmia Detectors". In: *Computers in Cardiology* (1984) (cit. on pp. 1, 168).

[132] V. Moskalenko, N. Zolotykh, and G. Osipov. "Deep Learning for ECG Segmentation". In: *Advances in Neural Computation, Machine Learning, and Cognitive Research III*. Ed. by B. Kryzhanovsky et al. Cham: Springer International Publishing, 2020, pp. 246–254. ISBN: 978-3-030-30425-6 (cit. on p. 32).

[133] M. Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015. ISBN: 978-3-319-21944-8 (cit. on pp. 18, 35, 36, 42, 48, 136).

[134] M. Müller and F. Zalkow. "FMP Notebooks: Educational Material for Teaching and Learning Fundamentals of Music Processing". In: *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*. Delft, The Netherlands, 2019-11 (cit. on pp. 18, 35, 36, 48).

[135] M. Müller and F. Zalkow. "libfmp: A Python Package for Fundamentals of Music Processing". In: *Journal of Open Source Software* 6.63 (2021), p. 3326. DOI: 10.21105/joss.03326. URL: https://doi.org/10.21105/joss.03326 (cit. on p. 47).

[136] P. K. Muthumanickam et al. "Shape grammar extraction for efficient query-by-sketch pattern matching in long time series". In: *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2016, pp. 121–130. DOI: 10.1109/VAST.2016.7883518 (cit. on p. 38).

[137] U. Neisser. *Cognitive Psychology: Classic Edition*. Psychology Press, 2014, pp. 282–290. ISBN: 1848726945 (cit. on p. 2).

[138] M. Nirmalan and P. M. Dark. "Broader applications of arterial pressure wave form analysis". In: *Continuing Education in Anaesthesia Critical Care and Pain* 14.6 (2014), pp. 285–290. DOI: 10.1093/bjaceaccp/mkt078. eprint: /oup/backfile/content_public/journal/ceaccp/14/6/10.1093/bjaceaccp/mkt078/2/mkt078.pdf. URL: +%20http://dx.doi.org/10.1093/bjaceaccp/mkt078 (cit. on p. 101).

[139] M. L. Nunes et al. "Posture Risk Assessment in an Automotive Assembly Line Using Inertial Sensors". In: *IEEE Access* 10 (2022), pp. 83221–83235. DOI: 10.1109/ACCESS.2022.3196473 (cit. on p. 125).

[140] N. Nunes, T. Araújo, and H. Gamboa. "Two-modes Cyclic Biosignal Clustering based on Time Series Analysis." In: 2011-01, pp. 257–264. DOI: 10.13140/2.1.1524.3048 (cit. on p. 34).

[141] M. Nyan, F. E. Tay, and E. Murugasu. "A wearable system for pre-impact fall detection". In: *Journal of biomechanics* 41.16 (2008), pp. 3475–3481 (cit. on p. 40).

[142] C. O'Connor. *Library to generate random strings from regular expressions.* URL: https://github.com/crdoconnor/xeger (cit. on p. 139).

[143] E. Occhipinti. "OCRA: a concise index for the assessment of exposure to repetitive movements of the upper limbs". In: *Ergonomics* 41.9 (1998), pp. 1290–1311 (cit. on p. 4).

[144] Y. Ohgi. "Microcomputer-based acceleration sensor device for sports biomechanics-stroke evaluation by using swimmer's wrist acceleration". In: *SENSORS, 2002 IEEE*. Vol. 1. IEEE. 2002, pp. 699–704 (cit. on pp. 1, 40).

[145] P. Ordóñez et al. "Using modified multivariate bag-of-words models to classify physiological data". In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2011, pp. 534–539. ISBN: 9780769544090. DOI: 10.1109/ICDMW.2011.174 (cit. on p. 37).

[146] S. Patel et al. "A review of wearable sensors and systems with application in rehabilitation". In: *Journal of neuroengineering and rehabilitation* 9.1 (2012), pp. 1–17 (cit. on p. 40).

[147] J. Paulus, M. Müller, and A. Klapuri. "Audio-based Music Structure Analysis". In: *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*. Utrecht, The Netherlands, 2010, pp. 625–636 (cit. on pp. 36, 42, 46).

[148] T. Pavlidis. "Linguistic Analysis of Waveforms". In: *Software Engineering - Computer and Information Sciences*. Ed. by J. T. Tou. Vol. 2. Academic Press, 1971, pp. 203–224 (cit. on p. 37).

[149] T. Pavlidis. "Waveform Segmentation Through Functional Approximation". In: *IEEE Transactions on Computers* C-22.7 (1973), pp. 689–697. ISSN: 0018-9340. DOI: 10.1109/TC.1973.5009136 (cit. on p. 37).

[150] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 23–25, 72).

[151] M. Perslev et al. "U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cit. on p. 33).

[152] C. Piccardi. "On parameter estimation of chaotic systems via symbolic time-series analysis". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 16.4 (2006), p. 043115. ISSN: 1054-1500. DOI: 10.1063/1.2372714. URL: http://aip.scitation.org/doi/10.1063/1.2372714 (cit. on p. 37).

[153] V. Rajagopalan et al. "Pattern identification in dynamical systems via symbolic time series analysis". In: *Pattern Recognition* 40.11 (2007-11), pp. 2897–2907. ISSN: 0031-3203. DOI: 10.1016/J.PATCOG.2007.03.007. URL: https://www.sciencedirect.com/science/article/pii/S003132030700115X (cit. on p. 37).

[154] S. Raschka. *Naive Bayes and text classification*. 2014-10. URL: https://sebastianraschka.com/Articles/2014_naive_bayes_1.html (cit. on pp. 23, 24).

[155] J. Rodrigues. *Fiber optic sensors for ergonomic analysis in automotive industry*. 2019. URL: https://drive.google.com/file/d/1EHSkt8eQVzLEJUyHJLyTR4JFs-xD7C5I/view?usp=sharing (cit. on p. 124).

[156] J. Rodrigues. *Wearable Technology - Review of Existing Concepts and Ideas for Wearable Technology*. 2019. URL: https://drive.google.com/file/d/1EHSkt8eQVzLEJUyHJLyTR4JFs-xD7C5I/view?usp=sharing (cit. on p. 124).

[157] J. Rodrigues et al. "Feature-Based Information Retrieval of Multimodal Biosignals with a Self-Similarity Matrix: Focus on Automatic Segmentation". In: *TBD* (2023), TBD (cit. on p. 135).

[158] J. Rodrigues, D. Belo, and H. Gamboa. "Noise detection on ECG based on agglomerative clustering of morphological features". In: *Computers in Biology and Medicine* 87 (2017), pp. 322–334. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2017.06.009. URL: http://www.sciencedirect.com/science/article/pii/S0010482517301737 (cit. on p. 40).

[159] J. Rodrigues, P. Probst, and H. Gamboa. "TSSummarize: A Visual Strategy to Summarize Biosignals". In: *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*. 2021, pp. 1–6. DOI: 10.1109/ICBSII51839.2021.9445154 (cit. on p. 135).

[160] J. Rodrigues et al. "A Genetic Algorithm to Design Job Rotation Schedules with Low Risk Exposure". In: *Technological Innovation for Life Improvement*. Ed. by L. M. Camarinha-Matos et al. Cham: Springer International Publishing, 2020, pp. 395–402. ISBN: 978-3-030-45124-0 (cit. on p. 4).

[161] J. Rodrigues et al. "microErgo: A Concept for an Ergonomic Self-Assessment Tool". In: *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*. 2021, pp. 1–6. DOI: 10.1109/ICBSII51839.2021.9445156 (cit. on pp. 125, 129).

[162] J. Rodrigues et al. "SSTS: A syntactic tool for pattern search on time series". In: *Information Processing and Management* 56.1 (2019), pp. 61–76. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2018.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0306457318302577 (cit. on pp. 35, 135).

[163] Y. Roh, G. Heo, and S. E. Whang. "A survey on data collection for machine learning: a big data-ai integration perspective". In: *IEEE Transactions on Knowledge and Data Engineering* (2019) (cit. on pp. 1, 2).

[164] D. Romero et al. "Towards an Operator 4.0 Typology: A Human-Centric Perspective on the Fourth Industrial Revolution Technologies". In: *International conference on computers and industrial engineering (CIE46) proceedings*. 2016-10 (cit. on p. 4).

[165] N. Roth et al. "Hidden Markov Model based Stride Segmentation on Unsupervised Free-living Gait Data in Parkinson's Disease Patients". In: (2021-02). DOI: 10.21203/rs.3.rs-269965/v1 (cit. on p. 33).

[166] S. Santos et al. "Explaining the Ergonomic Assessment of Human Movement in Industrial Contexts:" in: *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*. Valletta, Malta: SCITEPRESS - Science and Technology Publications, 2020, pp. 79–88. ISBN: 978-989-758-398-8. DOI: 10.5220/0008953800790088. URL: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0008953800790088 (visited on 2021-01-11) (cit. on pp. 40, 125, 171, 172).

[167] S. Santos et al. "Exploring Inertial Sensor Fusion Methods for Direct Ergonomic Assessments". In: *Biomedical Engineering Systems and Technologies*. Ed. by X. Ye et al. Cham: Springer International Publishing, 2021, pp. 289–303. ISBN: 978-3-030-72379-8 (cit. on pp. 1, 4, 125, 130, 172).

[168] A. Santos. et al. "Self-Similarity Matrix of Morphological Features for Motion Data Analysis in Manufacturing Scenarios". In: *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS*, INSTICC. SciTePress, 2021, pp. 80–90. ISBN: 978-989-758-490-9. DOI: 10.5220/0010252800800090 (cit. on pp. 1, 135).

[169] P. Schäfer. "The BOSS is Concerned with Time Series Classification in the Presence of Noise". In: *Data Min. Knowl. Discov.* 29.6 (2015-11), pp. 1505–1530. ISSN: 1384-5810. DOI: 10.1007/s10618-014-0377-7. URL: https://doi.org/10.1007/s10618-014-0377-7 (cit. on p. 39).

[170] P. Schäfer and M. Högqvist. "SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets". In: *EDBT '12*. 2012 (cit. on p. 39).

[171] P. Schäfer and U. Leser. "Fast and Accurate Time Series Classification with WEASEL". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 637–646. ISBN: 9781450349185. URL: https://doi.org/10.1145/3132847.3132980 (cit. on p. 39).

[172] K. Schaub et al. "The European Assembly Worksheet". In: *Theoretical Issues in Ergonomics Science* 14.6 (2013), pp. 616–639. DOI: 10.1080/1463922X.2012.678283. eprint: https://doi.org/10.1080/1463922X.2012.678283. URL: https://doi.org/10.1080/1463922X.2012.678283 (cit. on p. 4).

[173] S. Schneegaß et al. "A data set of real world driving to assess driver workload". In: New York, NY, USA: Association for Computing Machinery, 2013-10, pp. 150–157. ISBN: 9781450324786. DOI: 10.1145/2516540.2516561. URL: https://doi.org/10.1145/2516540.2516561 (cit. on pp. 120, 121, 170).

[174] P. Senin and S. Malinchik. "SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model". In: *2013 IEEE 13th International Conference on Data Mining*. 2013, pp. 1175–1180. DOI: 10.1109/ICDM.2013.52 (cit. on pp. 39, 40, 74).

[175] P. Senin et al. "Time series anomaly discovery with grammar-based compression." In: *EDBT*. 2015, pp. 481–492 (cit. on p. 37).

[176] G. Shakeri, J. H. Williamson, and S. Brewster. "Novel Multimodal Feedback Techniques for In-Car Mid-Air Gesture Interaction". In: AutomotiveUI '17. Oldenburg, Germany: Association for Computing Machinery, 2017, pp. 84–93. ISBN: 9781450351508. DOI: 10.1145/3122986.3123011. URL: https://doi.org/10.1145/3122986.3123011 (cit. on p. 115).

[177] H. Shatkay. "The Fourier Transform - a Primer". In: *Technical Report CS-95-37*. Brown University, 1995 (cit. on p. 11).

[178] L. Silva et al. "Respiratory Inductance Plethysmography to Assess Fatigability during Repetitive Work". In: *Sensors* 22 (2022-06), p. 4247. DOI: 10.3390/s22114247 (cit. on p. 4).

[179] P. Silva et al. "Towards better heartbeat segmentation with deep learning classification". In: *Scientific Reports* 10 (2020-11), p. 20701. DOI: 10.1038/s41598-020-77745-0 (cit. on p. 32).

[180] S. Silva. et al. "Assessing Occupational Health with a Cross-platform Application based on Self-reports and Biosignals". In: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF*, INSTICC. SciTePress, 2022, pp. 549–556. ISBN: 978-989-758-552-4. DOI: 10.5220/0010846700003123 (cit. on p. 125).

[181] M. Staudacher et al. "A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep". In: *Physica A: Statistical Mechanics and its Applications* 349.3 (2005), pp. 582–596. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2004.10.026. URL: http://www.sciencedirect.com/science/article/pii/S0378437104013640 (cit. on p. 1).

[182] M. Sung, C. Marci, and A. Pentland. "Wearable feedback systems for rehabilitation". In: *Journal of neuroengineering and rehabilitation* 2.1 (2005), pp. 1–12 (cit. on p. 40).

[183] H. Tankovska. *Global connected wearable devices 2016-2022*. 2020-09. URL: https://www.statista.com/statistics/487291/global-connected-wearable-devices/ (cit. on p. 1).

[184] E. Team. *The exponential growth of data*. 2018-04. URL: https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data (cit. on p. 1).

[185] G. Tirabassi and C. Masoller. "Unravelling the community structure of the climate system by using lags and symbolic time-series analysis". In: *Scientific Reports* 6 (2016). ISSN: 20452322. DOI: 10.1038/srep29804 (cit. on p. 37).

[186] P. Trahanias and E. Skordalakis. "Syntactic pattern recognition of the ECG". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.7 (1990-07), pp. 648–657. ISSN: 0162-8828. DOI: 10.1109/34.56207 (cit. on p. 37).

[187] C. Truong, L. Oudre, and N. Vayatis. "Selective review of offline change point detection methods". In: *Signal Processing* 167 (2020-02), p. 107299. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2019.107299. URL: http://dx.doi.org/10.1016/j.sigpro.2019.107299 (cit. on pp. 32–34, 86).

[188] J. K. Udupa and I. S. N. Murthy. "Syntactic Approach to ECG Rhythm Analysis". In: *IEEE Transactions on Biomedical Engineering* BME-27.7 (1980-07), pp. 370–375. ISSN: 0018-9294. DOI: 10.1109/TBME.1980.326650 (cit. on p. 37).

[189] A. S. Uva et al. *Lesões Musculoesqueléticas Relacionadas com o Trabalho. Guia de Orientação para a Prevenção*. 2008 (cit. on p. 4).

[190] R. Varandas., D. Folgado., and H. Gamboa. "Evaluation of Spatial-Temporal Anomalies in the Analysis of Human Movement". In: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4: BIOSIGNALS,* INSTICC. Prague, Czech Republic: SciTePress, 2019, pp. 163–170. ISBN: 978-989-758-353-7. DOI: 10.5220/0007386701630170. URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007386701630170 (visited on 2019-04-07) (cit. on pp. 4, 34, 40).

[191] P. Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cit. on p. 20).

[192] M. Wang et al. "Research on EMG segmentation algorithm and walking analysis based on signal envelope and integral electrical signal". In: *Photonic Network Communications* 37 (2019-04). DOI: 10.1007/s11107-018-0809-1 (cit. on p. 33).

[193] Q. Wang et al. "Unsupervised Temporal Segmentation of Repetitive Human Actions Based on Kinematic Modeling and Frequency Analysis". In: *Proceedings - 2015 International Conference on 3D Vision, 3DV 2015*. Institute of Electrical and Electronics Engineers Inc., 2015-11, pp. 562–570. ISBN: 9781467383325. DOI: 10.1109/3DV.2015.69. arXiv: 1512.04115 (cit. on p. 34).

[194] X. Wang et al. "Experimental Comparison of Representation Methods and Distance Measures for Time Series Data". In: *Data Mining and Knowledge Discovery* 26 (2010-12). DOI: 10.1007/s10618-012-0250-5 (cit. on p. 38).

[195] H. Wang., M. Mohamed Refai, and B. F. van Beijnum. "Measuring Upper-Extremity Use with One IMU". In: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4: BIOSIGNALS*, INSTICC. SciTePress, 2019, pp. 93–100. ISBN: 978-989-758-353-7. DOI: 10.5220/0007253400930100 (cit. on p. 4).

[196] M. Wattenberg. "Arc diagrams: visualizing structure in strings". In: *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.* 2002, pp. 110–116. DOI: 10.1109/INFVIS.2002.1173155 (cit. on pp. 36, 57).

[197] G. I. Webb. "Bayes Rule". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston, MA: Springer US, 2010, pp. 74–75. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_62. URL: https://doi.org/10.1007/978-0-387-30164-8_62 (cit. on pp. 23, 24).

[198] G. I. Webb. "Naïve Bayes". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston, MA: Springer US, 2010, pp. 713–714. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_576. URL: https://doi.org/10.1007/978-0-387-30164-8_576 (cit. on p. 24).

[199] A. Wege and A. Zimmermann. "Electromyography sensor based control for a hand exoskeleton". In: *2007 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2007, pp. 1470–1475 (cit. on p. 40).

[200] J. Weiner et al. "Bremen Big Data Challenge 2017: Predicting University Cafeteria Load". In: *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer. 2017, pp. 380–386 (cit. on p. 40).

[201] G. Weiss, K. Yoneda, and T. Hayajneh. "Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living". In: *IEEE Access* PP (2019-09), pp. 1–1. DOI: 10.1109/ACCESS.2019.2940729 (cit. on p. 166).

[202] A. Wilden. "The Rules Are No Game: The Strategy of Communication by Anthony Wilden". In: (1987), pp. 137–138 (cit. on p. 64).

[203] L. D. Xu, E. L. Xu, and L. Li. "Industry 4.0: state of the art and future trends". In: *International Journal of Production Research* 56.8 (2018), pp. 2941–2962 (cit. on p. 40).

[204] T. Xue and H. Liu. "Hidden Markov Model and Its Application in Human Activity Recognition and Fall Detection: A Review". In: *Communications, Signal Processing, and Systems*. Springer Singapore, 2022, pp. 863–869. ISBN: 978-981-19-0390-8. DOI: 10.1007/978-981-19-0390-8_108 (cit. on p. 40).

[205] E. Yakushenko et al. *St Petersburg INCART 12-lead arrhythmia database*. 2008-05. URL: https://physionet.org/content/incartdb/1.0.0/ (cit. on p. 169).

[206] P. Yang, G. Dumont, and J. M. Ansermino. "Adaptive Change Detection in Heart Rate Trend Monitoring in Anesthetized Children". In: *IEEE Transactions on Biomedical Engineering* 53.11 (2006), pp. 2211–2219. DOI: 10.1109/TBME.2006.877107 (cit. on p. 1).

[207] D. Yankov, E. Keogh, and S. Lonardi. "Dot plots for time series analysis". In: *17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)*. 2005, 10 pp.–168. DOI: 10.1109/ICTAI.2005.60 (cit. on p. 34).

[208] L. Ye and E. Keogh. "Time series shapelets: a new primitive for data mining". In: 2009-06, pp. 947–956. DOI: 10.1145/1557019.1557122 (cit. on p. 39).

[209] C. C. M. Yeh et al. "Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile". In: *Data Mining and Knowledge Discovery* 32.1 (2018-01), pp. 83–123. ISSN: 1573756X. DOI: 10.1007/s10618-017-0519-9. URL: https://doi.org/10.1007/s10618-017-0519-9 (cit. on p. 34).

[210] J. Yu et al. "Multivariate Multiscale Symbolic Entropy Analysis of Human Gait Signals". In: *Entropy* 19.10 (2017). ISSN: 1099-4300. DOI: 10.3390/e19100557. URL: http://www.mdpi.com/1099-4300/19/10/557 (cit. on p. 37).

[211] O. Yuji. "Mems sensor application for the motion analysis in sports science". In: *Memory* 32 (2005), 128Mbit (cit. on pp. 1, 40).

[212] Y. Zhang et al. "Multi-scale signed recurrence plot based time series classification using inception architectural networks". In: *Pattern Recognition* 123 (2022), p. 108385. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2021.108385. URL: https://www.sciencedirect.com/science/article/pii/S0031320321005653 (cit. on p. 34).

[213] C. Zhou et al. "Ankle foot motion recognition based on wireless wearable sEMG and acceleration sensors for smart AFO". In: *Sensors and Actuators A: Physical* 331 (2021), p. 113025 (cit. on p. 40).

[214] C. Zhou et al. "Research and Development of Ankle-Foot Orthoses: A Review". In: *Sensors* 22.17 (2022). ISSN: 1424-8220. DOI: 10.3390/s22176596 (cit. on p. 40).

A

Appendix 1 - Data Description and Management

In this work, we applied the proposed methodologies in a variety of publicly available datasets. Some datasets were used to exemplify how a proposed method could be valuable, in other cases, the datasets were used for validation. The methods developed and presented in this work are agnostic to the type of time series, but we focus all of the examples in *biosignals* data in order to demonstrate the applicability of such methods in the biomedical domain. In addition, most of the *biosignals* used are very common to be acquired in occupational scenarios. Most validations were performed with benchmark

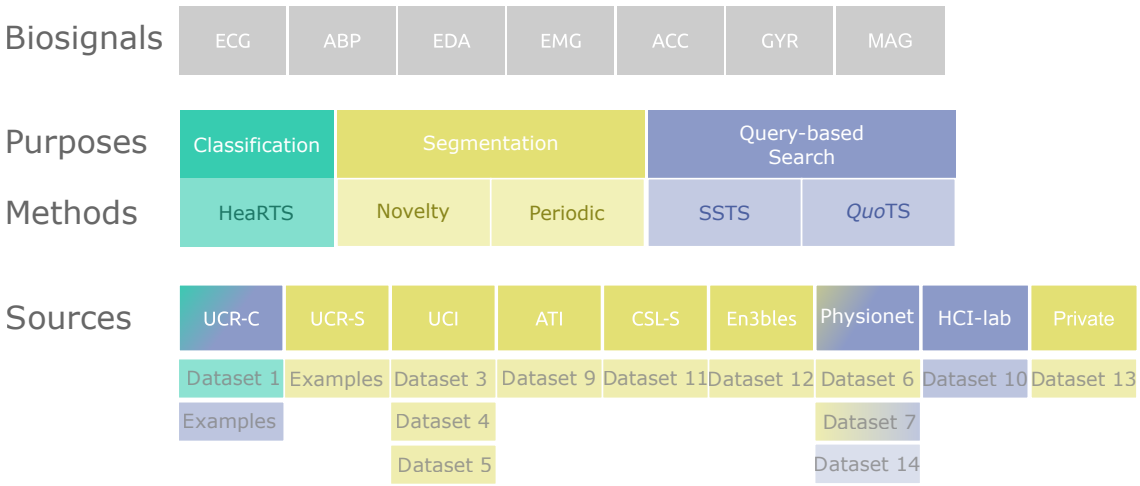


Figure A.1: Summary of the data used in this work. All types of biosignals where the methods developed in this work were applied to are listed. The three main purposes of the methods developed are presented and color-coded to highlight on each data source which methods were used. *UCR-C*: University California Riverside, and *C* stands for classification; *UCR-S*: *S* stands for segmentation; *UCI*: University California Irvine; *ATI:* Alan Turing Institute; *CSL-S:* Cognitive Systems Lab, where *S* stands for *share*; *HCI-lab:* Human Computer Interaction lab. The tag *Examples* means that several signals were used to demonstrate the usage of the methods as an example. The private source corresponds to the dataset recorded at *Volkswagen Autoeuropa*.

datasets, commonly used by the data science community for the tasks approached in this work, and have comparable measures from other existing methods.

Considering the diversity of topics covered, methods developed and datasets used, we summarized the information in Figure A.1. We show the types of *biosignals* that were used, the main sources for the publicly available data, and which their purpose in this work. From public datasets we used ECG, ABP, EMG, SCR, ACC and GYR biosignals. The purposes of these datasets involved mainly three tasks in the time series data mining domain: classification, segmentation and query-based search. We highlighted the colors corresponding to the purpose that the dataset would be use for.

A short description of each dataset is provided in this Section, highlighting their purpose in this work. In case an illustration of the signal is not provided during the explanation of the methods' or results' sections, we show an illustrative example of the signal to accompany the description of the dataset.

A.1 Datasets

A.1.1 Dataset 1 - *UCRC* - UCR Classification Benchmark

Description

The University of California Riverside (UCR) Time Series Archive was introduced in 2002 and is one of the most used benchmarks for time series data mining tasks, specially for classification. The datasets are diverse in terms of data type, data domain, difficulty, number of classes and dimensionality[44]. Several *python* distributions make it available to download. In this thesis, all UCR archive datasets from the *pyts* distribution were used for classification tasks. These represent 107 datasets from 18 different data types (audio, devices, ECG, Electrocularogram (EOG), EEG, HAR, etc...), which also are from many different domains, such as medical, financial, motion, entomology, etc...[44]. The list of datasets can be found on the following link [14].

Purpose

The UCR-C dataset is the benchmark used to validate HeaRTS for the classification task. Several signals from this dataset are used as examples as well to demonstrate the usage of QuoTS in query-based search tasks.

Ground Truth

The benchmark has a guideline in how to use it correctly at [44]. It has a pre-defined training and testing set. In case parameters have to by estimated during the training stage, it is recommended to use a cross-validation on the training set. In this work, these guidelines were followed to validate HeaRTS.

A.1.2 Dataset 2 - *TILT* - Physiological Response to Changes in Posture

Description

In the dataset [68], ten subjects' slow tilt, rapid tilt and standing-up activities were

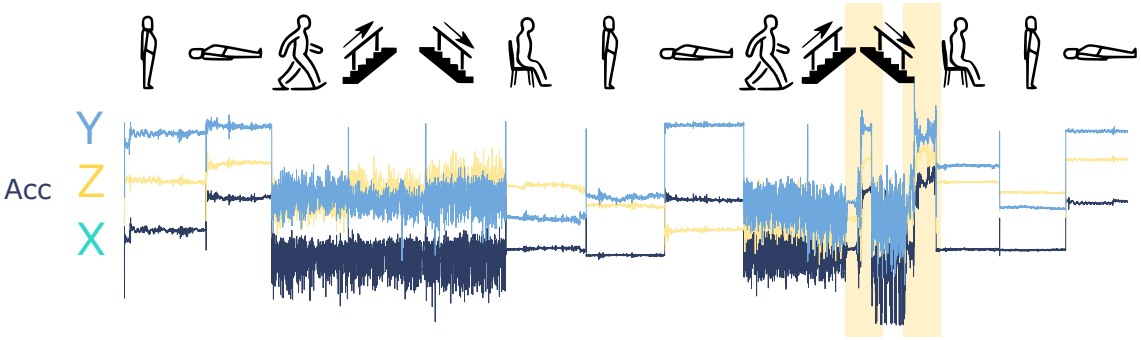


Figure A.2: Example of the signal for this dataset. It shows the 3 axis of the accelerometer signal and the corresponding labels in each section [10, 9]. Yellow areas indicate moments where there was a change on the signal not related with the labeled activity.

monitored and recorded with ECG and ABP to investigate how the two physiological signals respond to the angular changes during the mentioned activities [68, 60].

Purpose

The dataset's ABP channel was used as an example to demonstrate the usage of the SSM's novelty function in detecting pattern-based physiological changes in a distinctive research instance.

Ground Truth

The ground truth of the changes is marked by the angular signal, suggesting the moments a tilt or standing up activity occurred. The angular signal can have different shapes (rectangular or trapezoid), which depends on the speed at which the posture changes.

A.1.3 Dataset 3 - *HAR1* - Smartphone Dataset for Human Activity Recognition in Ambient Assisted Living

Description

This dataset was gathered from an experiment on 30 volunteers. Each subject was wearing a smartphone on the waist while performing several activities: *(1) Walking, (2) Walking Upstairs, (3) Walking Downstairs, (4) Sitting, (5) Standing and (6) Laying*. The activities were performed for approximately 60 seconds. The device recorded the internal ACC and GYR data at a constant rate of 50 Hz. Each activity has been categorized and labeled on the acquired data [10, 9].

Purpose

This dataset was used in the context of novelty segmentation [10, 9].

Ground Truth

Each file has a ground-truth signal that can be used to validate the novelty segmentation approach with the SSM's novelty function. A specific value is attributed to each class, which changes continuously over time and is aligned with the signal. The transitions between these activities are kept as the reference events for the validation.

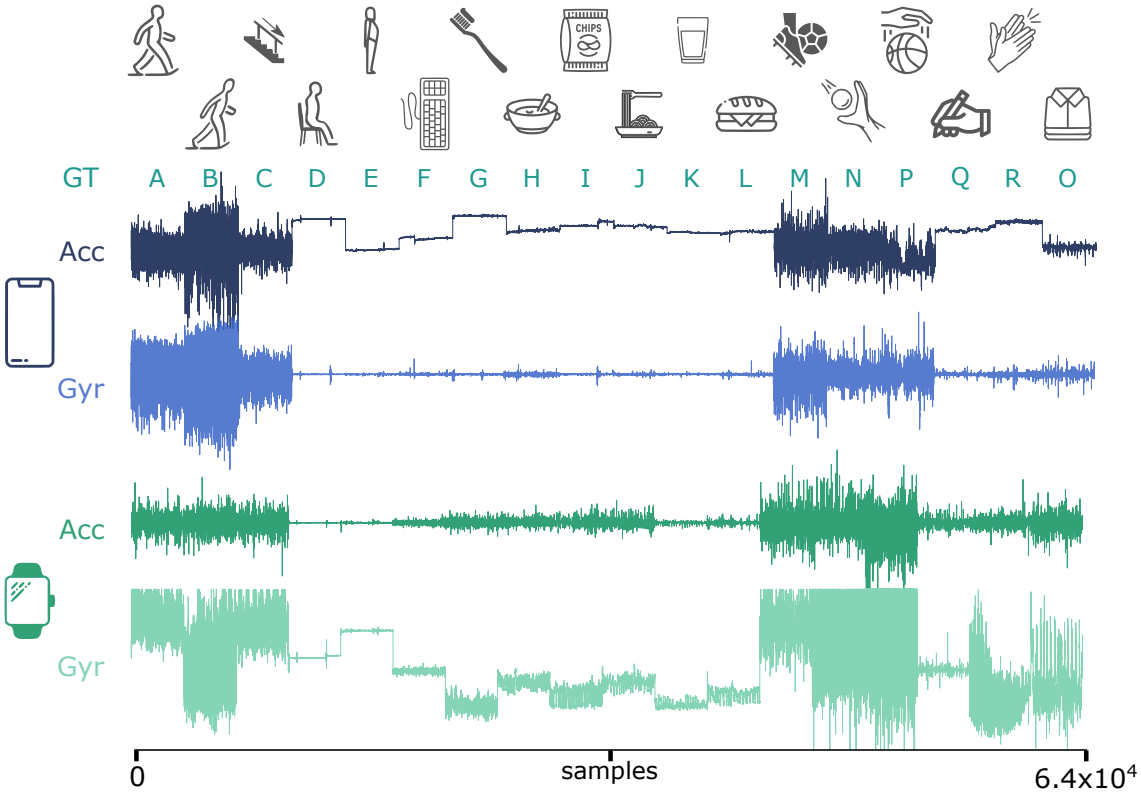


Figure A.3: Example of the signal for this dataset. It shows one axis of each type of data acquired, for both smartphone and smartwatch. The activities are highlighted as well and labeled as: A - walking, B - running, C - walking on stairs, D - sitting, E - standing, F - typing, G - brushing, H - eating soup, I - eating chips, J - eating pasta, K - drinking, L - eating a sandwich, M - kicking, N - catching a ball, P - dribbling, Q - writing, R - clapping, O - iron [201].

A.1.4 Dataset 4 - *WISDM* - Wireless Sensor Data Mining (WISDM) Smarphone and Smartwatch Activity Biometrics Dataset

Description

The raw data from the accelerometer and gyroscope sensors is collected from a smartphone and smartwatch at a rate of 20Hz. This experiment was conducted on 51 participants as they performed 18 activities, each for a duration of 3 minutes. Each sample of the data was labelled based on the activity it corresponds to. The activities are diverse and include dribbling, eating, jogging, sitting, walking on stairs, standing, walking, among others [201]. Figure A.3 illustrates an example of this type of data.

Purpose

This dataset was used in the context of novelty segmentation in complex real-life scenarios.

Ground Truth

The ground truth is available for each file and each activity. For the task of novelty search, we used the transition between labels as the reference instant of a change, as we did with the previous dataset.

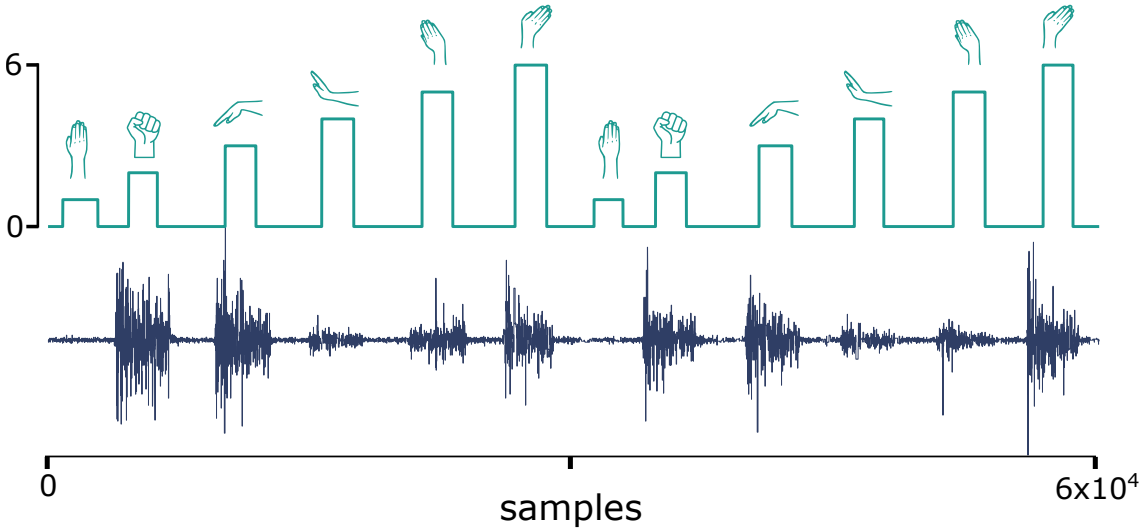


Figure A.4: Example of the EMG data and the corresponding hand postures.

A.1.5 Dataset 5 - *EMG* - EMG Data for Gestures

Description

This dataset has EMG signals for recording patterns, by using a MYO Thalmic bracelet worn on a user's forearm. The bracelet is equipped with eight sensors equally spaced around the forearm that simultaneously acquire electromyographic signals. The dataset has raw EMG data from 36 subjects while they performed series of static hand gestures. The subject performs two series, each of which consists of six basic gestures. Each gesture was performed for 3 seconds with a pause of 3 seconds between gestures. The data was collected with a fixed sampling frequency of 200 Hz [111].

Purpose

This dataset was used in the context of novelty segmentation, to test the method in estimating transitions between the activation (onset) and relaxation (offset) of the muscular activity.

Ground Truth

The ground truth is available for each file and each activity. For the task of novelty search, we used the transition between labels as the reference instant of a change. As indicated in Figure A.4, the ground truth would have considered as events the absence of activity (which the original dataset uses because it was designed for a classification problem). In this case, we did not considered these events in our ground truth for this dataset for the problem of segmentation.

A.1.6 Dataset 6 - *ECG1* - MIT-BIH Noise Stress Test Database

Description

The dataset comprehends 12 half-hour ECG recordings and 3 half-hour recordings of noise typical in ambulatory ECG recordings. The noise recordings were made using

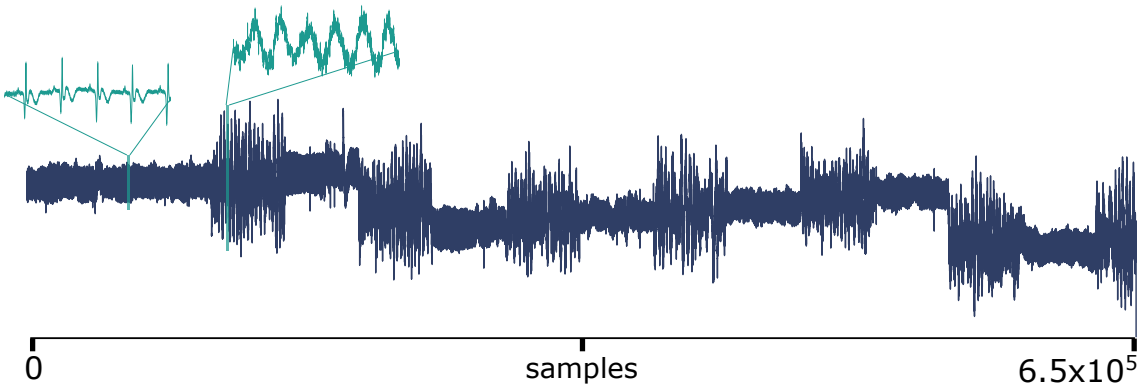


Figure A.5: Example of an ECG record contaminated with noise. Two sections are highlighted showing the clean and contaminated area of the signal. [19]

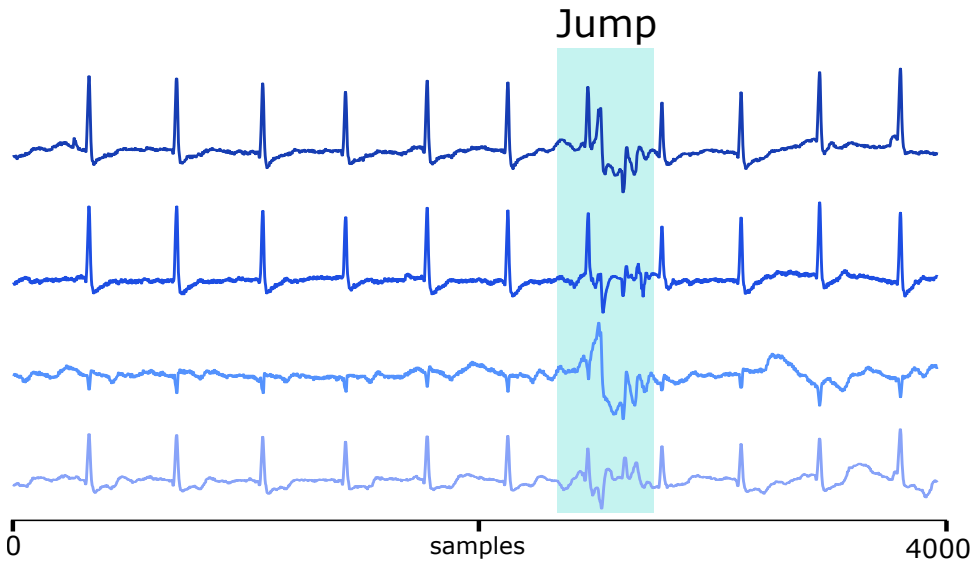


Figure A.6: Example of an ECG signal contaminated with motion artifacts. The artifact is visible due to a jump from the subject.

physically active volunteers and standard ECG recorders, leads, and electrodes. The three noise records were assembled from the recordings by selecting intervals that contained predominantly baseline wander (in record 'bw'), muscle (EMG) artifact (in record 'ma'), and electrode motion artifact (in record 'em'). Two clean ECG signals were selected and noise was added with different signal-to-noise ratios (SNR) [131, 60].

Purpose

This dataset was used in the context of novelty segmentation, to test the method in estimating transitions to and from noise sections of the signal.

Ground Truth

The reference events were manually annotated based on the signal with lower SNR. All noise presence is located in the same regions of the signal, only with different SNR.

A.1.7 Dataset 7 - *ECG2* - Motion Artifacted Contaminated ECG

Description

This dataset has short duration ECG signals, which were recorded from a healthy 25-year-old male performing different physical activities (standing, walking and single jump) to study the effect of motion artifacts on ECG signals and their sparsity. The dataset was acquired with a sampling rate of 500 Hz and 16 bit resolution. For this exercise, only the records with jump were used [19, 60].

Purpose

This dataset was used in the context of change point detection, to test the method in estimating transitions to and from noise sections of the signal.

Ground Truth

The reference events were manually annotated in the intervals of noise due to a jump activity.

A.1.8 Dataset 8 - *INCART* - St Petersburg INCART 12-lead Arrhythmia Database

Description

This dataset has ECG signals from patients undergoing tests for coronary artery disease. 75 annotated recordings were extracted. Each record is 30 minutes long and has 12 leads, each sampled at 257 Hz. The annotations include beats as well as relevant artifacts/occurrences present on the signal, such as arrhythmia [60]. Records *I02, I04 and I09* were used.

Purpose

This dataset was used in the context of pattern search with QuoTS for the detection of different types of arrythmia.

Ground Truth

The ECG beats were detected by an automatic algorithm, which were then manually corrected when needed. The location and type of arrhythmia is available on the dataset [205].

A.1.9 Dataset 9 - *ATCPD* - Alan Turing CPD Benchmark

Description

The Alan Turing Change Point Detection Benchmark is a recent dataset used to have a standard reference for change point segmentation tasks. The benchmark was created in 2020 and provides 42 datasets, being 42 univariate time series and 4 multi dimensional time series. The dataset comprises several time series from real-world data. It was built from a project of the Alan Turing Institute to have a repository for the evaluation of change point detection algorithms. The available benchmark page was used to get access to the datasets, and run the performance metrics developed. This way, the performance of the proposed method is compare with the performance of several existing methods [26].

Purpose

This dataset has ground-truth events for each time series. The performance of several existing methods is available and used to compare with the performance of the proposed method on the same time series.

Ground Truth

The signals available in this dataset were manually annotated by several human annotators. Each signal has, in most cases, more than one reference annotation. The evaluation of an algorithm's performance in event detection follows the same metrics as the ones mentioned in 2.8.2, but combines the information of multiple annotators [26].

A.1.10 Dataset 10 - *HCILAB* - HCILab Behavioural Driving Dataset

Description

It is a public dataset that studied ways of assessing the driver's workload combining auto telemetry data with physiological sensors. The authors conducted a real world driving experiment with 10 participants measuring a variety of physiological data as well as a post-hoc video rating session [173]. It includes a variety of physiological signals, such as ECG and corresponding heart rate, SCR and body temperature. In addition, it collected driving speed and GPS location. The map of the driving site can be seen at [173]

Purpose

This dataset was used for QuoTS in pattern search. For the example presented we used the data of participant number 10 [173].

Ground Truth

The dataset is not originally labelled. In this work, the signals were explored with the QuoTS tool to search for relevant patterns that could be associated with relevant occurrences during the driving experience. Although the videos were not available, the GPS data was provided and could be used to match the position of the car with the physiological signals.

A.1.11 Dataset 11 - *CSL* - CSL-Share Dataset

Description

The CSL-SHARE (Cognitive Systems Lab Sensor-based Human Activity REcordings) has twenty two types of activities of daily living and sports recorded in tewnty subjects. It contains multimodal data with two triaxial ACC, two triaxial GYR, four surface EMG sensors, one biaxial Goniometer (GON), and one airborne microphone integrated into a knee bandage [106]. Seventeen activities were used for this purpose, which were the following: *Sit*, *Sit-to-Stand*, *Stand*, *Stand-to-Sit*, *Stair-Up*, *Stair-Down*, *Walk*, *Curve-Left-Step*, *Curve-Left-Spin*, *Curve-Right-Step*, *Curve-Right-Spin*, *Run*, *V-Cut-Left*, *V-Cut-Right*, *Lateral-Shuffle-Left*, *Lateral-Shuffle-Right*, *Jump-One-Leg*, *Jump-Two-Leg*.

Purpose

Each activity was performed multiple times in one record. This dataset was used for the

onset periodic segmentation of each activity for each record of each subject.

Ground Truth

A ground-truth squared signal was generated with a push-button during the acquisitions [106]. The transitions between null and positive square wave indicate the onset and offset of the activity of a record. These transitions were used as the ground-truth.

A.1.12 Dataset 12 - *EN3BLES* - ENABL3S

Description

The ENABL3S dataset contains bilateral EMG and motion data from inertial wearable sensors positioned on joints and limbs. Ten subjects participated in the recording, which involved freely transitioning between sitting, standing, and five walking-related activities. In total, 23 sensors were used (14 EMG, 4 GON, 5 IMU) and 52 channels (14 EMG, 8 GON, 30 IMU - 5 triaxial ACC + 5 triaxial GYR) [76].

Purpose

The dataset was used for the segmentation of activities, made by finding the break-points on the multimodal time series.

Ground Truth

The labels of each activity were available for each dataset. The segmentation points used as ground truth are the transitions between labels.

A.1.13 Dataset 13 - *VAOD* - Volkswagen Autoeuropa Occupational Dataset

Description

A private dataset was acquired at *Volkswagen Autoeuropa* in the context of this thesis, a master thesis from [166] and project *OPERATOR*, in partnership with Fraunhofer AICOS. This dataset used a wearable system with inertial sensors to record direct measures that could be used to calculate the occupational exposure to risk factors of workers during work. The technologycal setting consisted of four 9-Domain of Freedom (DoF) IMU sensors (composed internally by a triaxial ACC, triaxial GYR and triaxial Magnetometer (MAG)) and an Android wireless communication system, which was responsible to record and allocate the data [166].

The acquisition was made in a *Volkswagen Industrial* assembly line where twelve manufacturing collaborators performed their tasks while having attached the system in their dominant upper body segment (back of hand, wrist, elbow and trunk). The data was recorded in three different workstations from the *Bodyshop assembly line*, which is one section where the cars' doors are assembled. The three workstations are designated: 1) Liftgate workstation, where back doors are mounted; 2) Fender workstation, involving front door tasks and 3) Doors workstation, which demanded tasks on the front doors and in the cars' hood [166]. A total of six operators performed at least two different workstations. All recordings were filmed for posterior visual evaluation. The videos and signals were synchronized by a simultaneous event visible in both video and inertial

signals. The event corresponds to a sequence of poses: 1) neutral anatomic position, 2) T pose, 3) return to neutral anatomic position, to finally 4) jump.

The mentioned study was centered on the ergonomic assessment of the dominant arm. For this reason, the IMUs were attached in: 1) the posterior side of the hand, 2) posterior side of the forearm and 3) posterior side of the arm and a final one 4) placed in the anterior side of the thorax area. All of the devices were attached with elastic bands, in such a way that all had their Y-axis pointed up while in a neutral anatomical position [166]. Additionally, a Smartphone was attached to the trunk of the collaborators as well, working as an additional IMU, from which the position of the arm in regards to the body posture was calculated.

Some important notes regarding exceptions found in sensors:

1. Despite the actual acquisition having three sensors ACC, GYR and MAG, only the ACC and GYR sensors were considered for this study as this had the best behaviour, and the MAG was acting in an erratic manner, probably due to the heavily magnetic environment during the acquisition.

2. The two workstations of operator 2 were made during the same acquisition, resulting in a single time series, where the subject performed two different types of active work motions.

3. In operator 2 workstations 1& 2, the torso data was not considered, due to malfunction.

More details regarding the dataset and the acquisition process can be found at [166, 167].

Purpose

This database was used to study the application of the algorithm to detect (1) Working Periods with the novelty function, (2) Periodic Working Cycles with the similarity function and (3) Search by example.

Ground Truth

The data was annotated based on video inspection. For this, first, the data of each individual sensors had to be aligned based on the reference event visible in all signals (T pose and jump). Afterwards, the annotations on the signals were made based on the events recorded on the video considering the common synchronous event.

A.1.14 Dataset 14 - *UCRS* - UCR Semantic Segmentation Benchmark

Description

The *UCR Semantic Segmentation Benchmark* [57] was recently designed with the intent to test segmentation performances for cases where there is a change in *regime* on time series. The dataset has several types of signals, with different difficulties. We used a few signals from this dataset, namely the *pulsus paradoxus* example. The signal represents an ECG

recorded from a patient that had an onset of *pulsus paradoxus* [35, 130].

Purpose

The signal was used as an illustrative application scenario for the proposed method.

Ground Truth

The signal has a regime change at the $10,000^{\text{th}}$ sample.

A.2 Notes for Ground Truth Annotations on Novelty Segmentation

Regarding the proposed method for novelty segmentation, the datasets used were mostly designed for classification tasks. Therefore, these were labeled with categorical values for each sample of the time series. As the problem of novelty segmentation requires the detection of specific samples of the signal, we adapted the labels of the datasets to fit the purpose of segmentation. This was made by only keeping the transitions between categories of labels, for example, in Figure **??**, the ground-truth (GT) is categorical. For evaluation purposes, the ground-truth was converted to a binary signal, where categorical transitions were valued as **1**, to keep only the position of the change, and not the category. This is valid for segmentation purposes only.

B

Appendix 2 - Detailed Results for Information Retrieval and Classification

B.1 TSFEL Feature List of Information Retrieval Experiments

The set of features extracted with the TSFEL *Python* library to compute the SSM are listed in Table B.1. These are not the entire set of features available, but the ones that could be optimize to increase the processing speed.

Table B.1: Features applied in this work to create the SSMs. The Time Series Feature Extraction Library (TSFEL) is used for feature extraction.

Temporal domain	**Statistical domain**	**Frequency domain**
Absolute energy	Interquartile Range	Entropy
Area under the curve	Kurtosis	Fundamental frequency
Centroid	Maximum	Max frequency
Cumulative centroid	Mean	Roll off
Distance	Mean absolute deviation	Roll on
Maximum peak	Median	Spectral distance
Mean absolute difference	Minimum	Spectral kurtosis
Mean difference	Root mean square	Spectral skewness
Median absolute difference	Skewness	Spectral spread
Total energy	Standard deviation	
	Variance	

B.2 Novelty Segmentation Detailed Results

The parameters used for the novelty search on each of the datasets where the method was applied are presented in the online repository [1] with excel files. Each table has the

[1] https://github.com/JmdRodrigues/PhDThesis/tree/main/ProjectCode/NOVA/Results/Tables

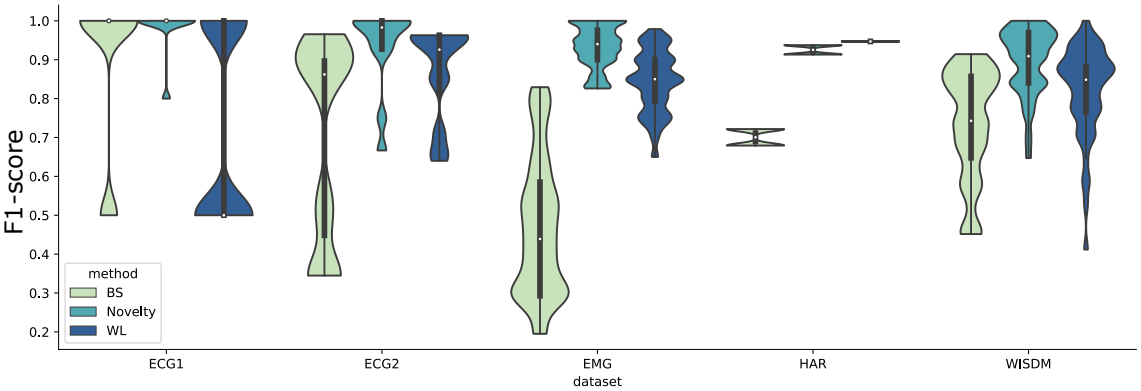


Figure B.1: F1-scores distribution for each method in each dataset. The X-axis indicates the dataset, the Y-axis the F1-scores. The labels of the methods are indicated on the bottom left.

Table B.2: Parameter configuration and experimental results of each signal in Dataset 3 (see Section A.1.3). W_{size}: window size; K: kernel size in ratio of the window size; O: overlap ratio of the window size. T: amplitude threshold ratio for the event detection; Novelty: our proposed method with the novelty function; WL: window-based segmentation; BS: binary segmentation. The tolerance used to calculate the F1-score ($F1$) equals the window size for novelty function computation.

	Novelty					WL		BS
Signal	W_{size}	K	O	T	$F1$	W_{size}	$F1$	$F1$
1	1500	1.5	0.90	0.01	0.91	1500	0.95	0.68
2	1500	1.5	0.90	0.01	0.94	1500	0.95	0.72

window size (W_{size}), the kernel size in ratio of the window size (K), the overlap ratio of the window size (O), and the amplitude threshold as the ratio of the maximum amplitude of the novelty function (T) for the detection of its positive peaks.

B.2.1 Scores Distribution

In Figure B.1 we present the overall distribution of F1-scores obtained for each record. These Figures are used to summarize the results available on the online repository, and complement the information provided on the final results.

The Figure presents information that corroborates the summary presented in the results section, where the F1-score is typically better for the proposed method, indicating that it is more robust than WL and BS methods for the segmentation task in these datasets.

Table B.6: Parameter configuration and experimental results of each signal in Dataset 5 (see Section A.1.4). W_{size}: window size; K: kernel size in ratio of the window size; O: overlap ratio of the window size. T: amplitude threshold for event detection; Novelty: our proposed method with the novelty function; WL: window-based segmentation; BS: binary segmentation. The tolerance used to calculate the F1-score ($F1$) equals the window size for novelty function computation.

	Novelty					WL		BS
Signal	W_{size}	K	O	T	$F1$	W_{size}	$F1$	$F1$
0	500	2.60	0.10	0.95	1.00	3000	0.85	0.71
1	500	6.50	0.14	0.95	0.91	1000	0.82	0.82
2	500	6.50	0.10	0.95	0.94	3600	0.76	0.82
3	1000	6.25	0.10	0.95	0.84	1000	0.76	0.59
4	500	6.50	0.10	0.95	0.90	2500	0.82	0.82
5	500	2.60	0.10	0.95	0.97	4000	0.97	0.88
6	500	2.60	0.14	0.95	0.97	3000	0.94	0.88
7	500	6.50	0.19	0.95	0.97	2500	0.94	0.81
8	500	6.50	0.10	0.95	1.00	3000	0.88	0.82
9	500	6.50	0.14	0.95	0.86	3000	0.71	0.69
10	500	6.50	0.10	0.95	0.97	3000	0.88	0.82
11	500	1.30	0.10	0.95	0.97	2500	1.00	0.88
12	500	6.50	0.10	0.95	0.94	2500	0.88	0.88
13	500	6.50	0.10	0.95	0.91	4000	0.91	0.88
14	500	2.60	0.10	0.95	0.97	3600	0.88	0.91
15	500	6.50	0.14	0.95	0.90	3000	0.79	0.86
16	1000	2.50	0.10	0.95	0.83	2500	0.87	0.45
17	500	2.60	0.14	0.95	1.00	2500	0.94	0.86
18	500	6.50	0.10	0.95	0.83	2500	0.88	0.91
19	500	2.60	0.14	0.95	0.84	4000	0.85	0.86
20	500	6.50	0.19	0.95	0.68	3000	0.71	0.57
21	500	6.50	0.19	0.95	0.84	3600	0.79	0.74
22	500	6.50	0.10	0.95	0.76	3000	0.94	0.74
23	500	2.60	0.10	0.95	0.97	3000	0.88	0.46
24	500	2.60	0.10	0.95	0.84	4000	0.84	0.69
25	500	6.50	0.10	0.95	0.91	4000	0.76	0.46
26	500	1.30	0.14	0.95	0.94	2500	0.88	0.91
27	500	2.60	0.41	0.95	0.76	4000	0.88	0.51
28	500	6.50	0.10	0.95	0.89	2500	0.71	0.57
29	500	6.50	0.10	0.95	1.00	3600	0.93	0.68

Continuation of Table B.6

	Novelty					WL		BS
Signal	W_{size}	K	O	T	$F1$	W_{size}	$F1$	$F1$
30	500	6.50	0.14	0.95	0.94	3600	0.91	0.86
31	500	6.50	0.10	0.95	0.97	2500	0.94	0.86
32	500	6.50	0.14	0.95	1.00	3600	0.87	0.86
33	500	6.50	0.10	0.95	0.84	3000	0.82	0.80
34	500	6.50	0.14	0.95	0.97	2500	0.94	0.86
35	500	6.50	0.10	0.95	0.94	4000	0.76	0.46
36	500	6.50	0.10	0.95	0.97	2500	0.87	0.74
37	500	6.50	0.10	0.95	1.00	2500	0.94	0.74
38	500	2.60	0.19	0.95	0.94	3600	0.87	0.74
39	500	6.50	0.14	0.95	0.86	3600	0.77	0.69
40	1000	6.25	0.10	0.95	0.94	4000	0.77	0.91
41	500	6.50	0.14	0.95	0.80	3600	0.59	0.57
42	500	6.50	0.10	0.95	0.87	3600	0.80	0.58
43	1000	2.50	0.10	0.95	0.91	3000	0.88	0.67
44	500	6.50	0.32	0.95	0.65	1000	0.41	0.46
45	1000	2.50	0.19	0.95	0.80	1000	0.53	0.63
46	500	2.60	0.32	0.95	0.79	3000	0.59	0.80
47	500	6.50	0.19	0.95	0.84	3600	0.71	0.69
48	500	6.50	0.10	0.95	0.91	3600	0.71	0.74
49	500	6.50	0.10	0.95	0.94	4000	0.85	0.74
50	500	6.50	0.19	0.95	0.71	3600	0.65	0.63

End of Table

B.2.2 A Few Notes on Event Distances

In Figure B.2 we show the distributions of distance errors for true positives found by each method in each dataset. The errors are presented in absolute samples and seconds. There is an obvious tendency for higher errors for the proposed method, indicating that it is less precise in identifying when the change happens, but still more reliable in identifying the overall set of changes that occur in the signal. All methods have the same increasing or decreasing error tendency, which can be associated with several factors. Cases where the errors are higher can be found for Datasets *HAR1*, *WISDM* and *ECG2*.

We believe there are several factors that influence the error or precision at which a method can find the changes accurately. These include the sampling frequency of the signal, the timescale at which the events are searched and happen, the labelling errors, and the dimension of the signal.

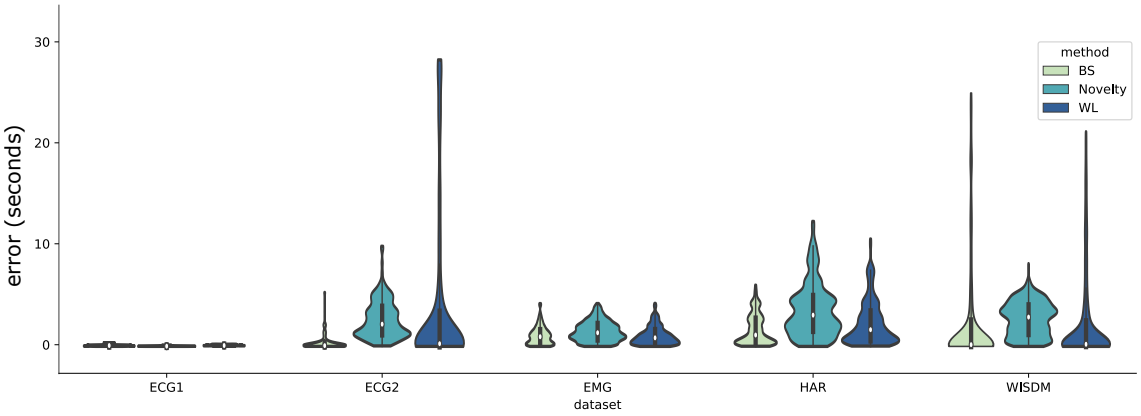


Figure B.2: Distance values of all *TP* identified by each method for all datasets. The X-axis indicates the dataset, while the Y-axis indicates the error in seconds as the ratio between the distance in samples and the sampling frequency.

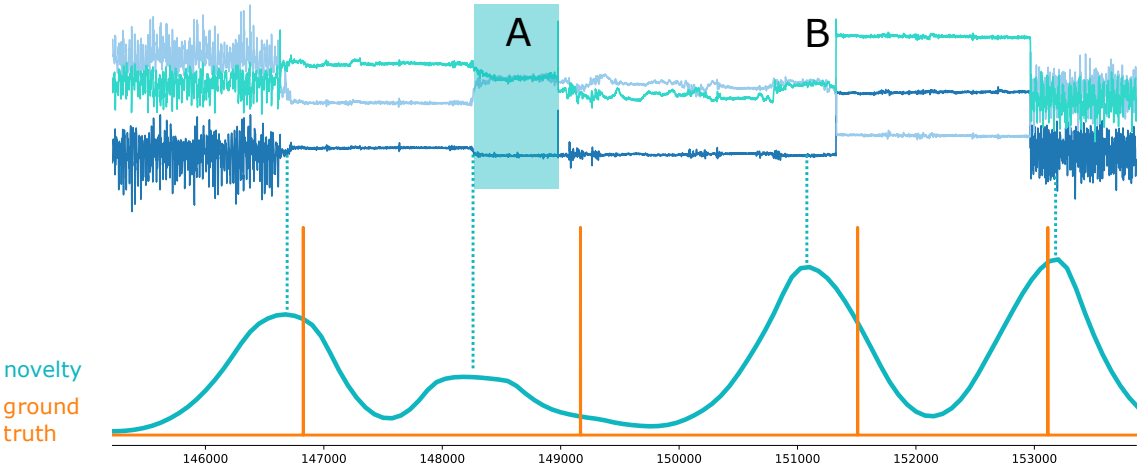


Figure B.3: Examples of imprecise labelling from Dataset *HAR1*. The 3-axis of the ACC signal are presented as well as the ground-truth (in orange) and the novelty function (in light blue), computed with the parameters used for the results obtained. Area **A** indicates the transition from standing to laying, while the transition **B** indicates the transition from laying to sitting.

A higher sampling frequency with a low timescale segmentation search makes the process less erroneous. In case the task is made on a very large signal (which is the case for *ECG2* signals), with a low sampling frequency and is searched in a very large timescale (several minutes), than the errors are prone to be much higher. This is the case for datasets *ECG2* and *WISDM*, where higher errors can be accounted by the larger search scale (several minutes in both cases) and large signals with low sampling frequency (20 Hz in case of *WISDM*). We believe this error is reasonable in case the user is exploring the data, but not in cases the user needs very accurate segmentations.

In case of *HAR1*, the increased error can also be explained by the lack of precision in some ground truth labels. As we show in Figure B.3, there are ground-truths that do not match precisely where the change happens and this has an effect on the distances. An

example is the peak that identifies the transition between sitting and standing (noted as **B** in Figure B.3). In this case, the distance was very high (approximately 500 samples, which accounts for 10 seconds), also because the ground-truth was far from the actual segmentation point. We believe it would be unfair to not count it as a *TP*, but we have to understand the limitations of such imprecise measurement. In addition, another example of imprecise labelling is presented in the same Figure (noted as A), where the change between standing and laying is not sharp and the transition takes some time. The proposed method identifies this change at the beginning of the transition, but the label is made at the end. This was counted as a *FP*, which is unfair. Still, we have to take these limitations into account and reason about how to improve both the method, but also, how to be fair in the evaluation of such segmentation tasks.

Typically, both *WL* and *BS* have better results in this aspect, but are not as robust in identifying all segmentation changes, which gives value to the proposed method.

How can we improve the method in being more precise? The method can be improved if a higher overlap is used. However this also implies that certain conditions are met, namely that the record is not too large, since the matrix might not be able to be entirely computed. Larger records require larger search windows, lower overlap and larger kernel sizes, which reduce the ability to detect short timescale events and make it less precise. One way to counteract this effect is to not compute the entire matrix, but rather only follow its diagonal while sliding the window and extracting the features. In this case, the method could be used with total overlap and the novelty function could be computed with higher precision. In addition, we also believe the detection strategy (using a peak finder) can make it less accurate, since we need a very smooth novelty function for this purpose. One possible strategy could involve the search of higher peaks in non overlaping windows, removing all other peaks that could be counted as *FP*.

B.3 Novelty Segmentation of Occupational Data

The novelty segmentation and periodic segmentation was performed on the occupational dataset recorded at *Volkswagen Autoeuropa*. The presented results show in Table B.8 the detection of changes that correspond to active to non-active working periods as well as transitions between two different workstations. All these events were annotated manually with the support of the videos available.

In Table B.9, the detection of each working period was made with the similarity function. As exemplified in other datasets, the valleys of this function indicate the presence of a period. The novelty function was used to segment the signal in smaller subsequences. For each of these segments, the number of valleys detected was counted, being indicated how many of the working cycles were detected in comparison with the real number of cycles. The criteria for this evaluation are the following:

1. **correct detection of cycles**. A cycle segment is considered correct if the moments

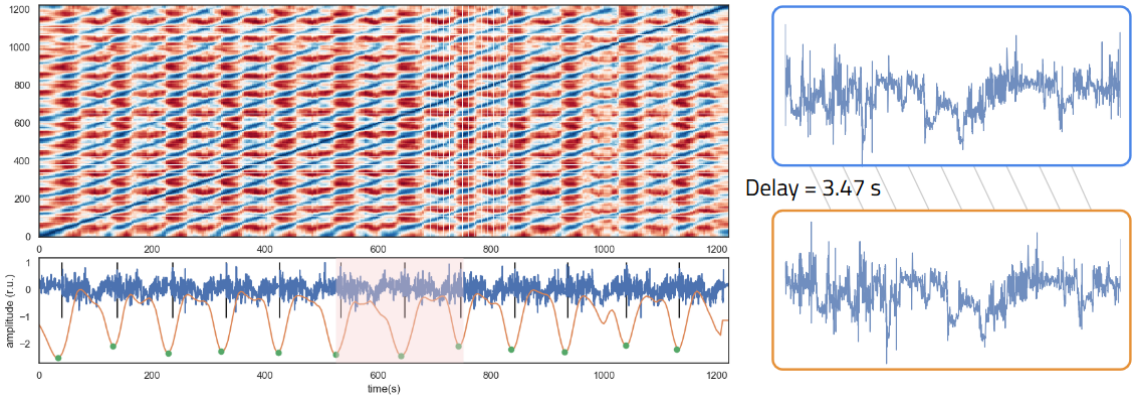


Figure B.4: Example of the detection of working cycles by means of the SSM after removing the non-active working periods. Black bars represent the ground truth events, green circles are the detected cycles on the similarity function (orange), represented on the bottom subplot. The X-accelerometer signal from the hand sensor is in blue.

at which the cycles are segmented correspond to a consistent position on the signal. Even if the segmentation of cycles occurs delayed from the ground-truth selection, what is evaluated in this category is the consistency of the algorithm in defining the working cycle. Here are measured how many cycles are correctly segmented;

2. **calculate the error between the ground-truth segmentation and the algorithm's segmentation**. The ground-truth duration of cycles is compared with the duration of the detected cycles by calculating the absolute difference between durations. This error is expressed in terms of seconds per cycle and duration percentage of the cycle. Figure B.4 shows an example of the detection performed, being the duration error how many seconds the interval between dark bars and green circles highlighted on the Figure differ. The percentage indicates the ratio of the working cycle that the duration error corresponds to.

B.4 Overall Results of HeaRTS and 1-NN ED on Dataset 1

B.4.1 Additional Results

In Figure B.5 are displayed the comparison between each text-based method and the 1-NN ED. These results show that the methods have a comparable performance with this standard approach. It also shows that the BoW or TF-idf models used with an SVM classifier performed better and tend to have better results. However, additional results should be performed to optimize the intrinsic hyperparameters of both SVM and NB models.

One relevant difference that was noticed with the usage of the text-based methods (HeaRTS), was the decrease in performance when applied on the validation step. Figure B.6 shows how this is evident for HeaRTS. The performance dropped substantially,

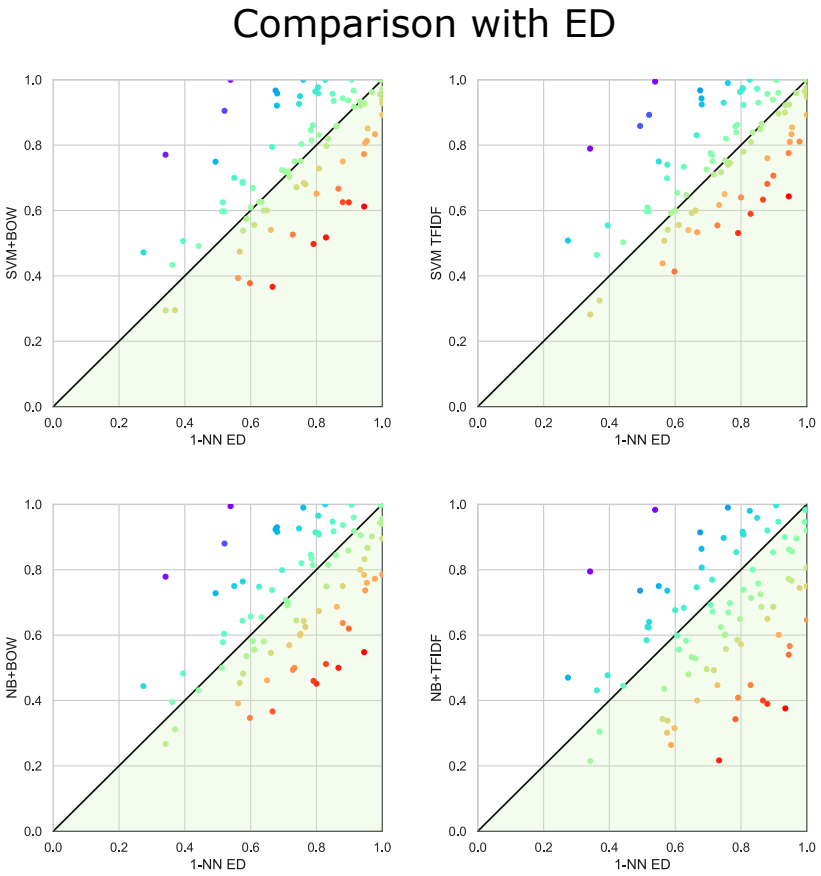


Figure B.5: Maps with the comparison between each text-based method explored and the 1-NN ED.

suggesting that the text models might not generalize as easily in some cases. The 1-NN ED method had a more constant performance.

B.4.2 Accuracies for each dataset

In table B.10 are displayed the overall results of all algorithms used on Dataset 1 (*UCRC*, see Section **??**).

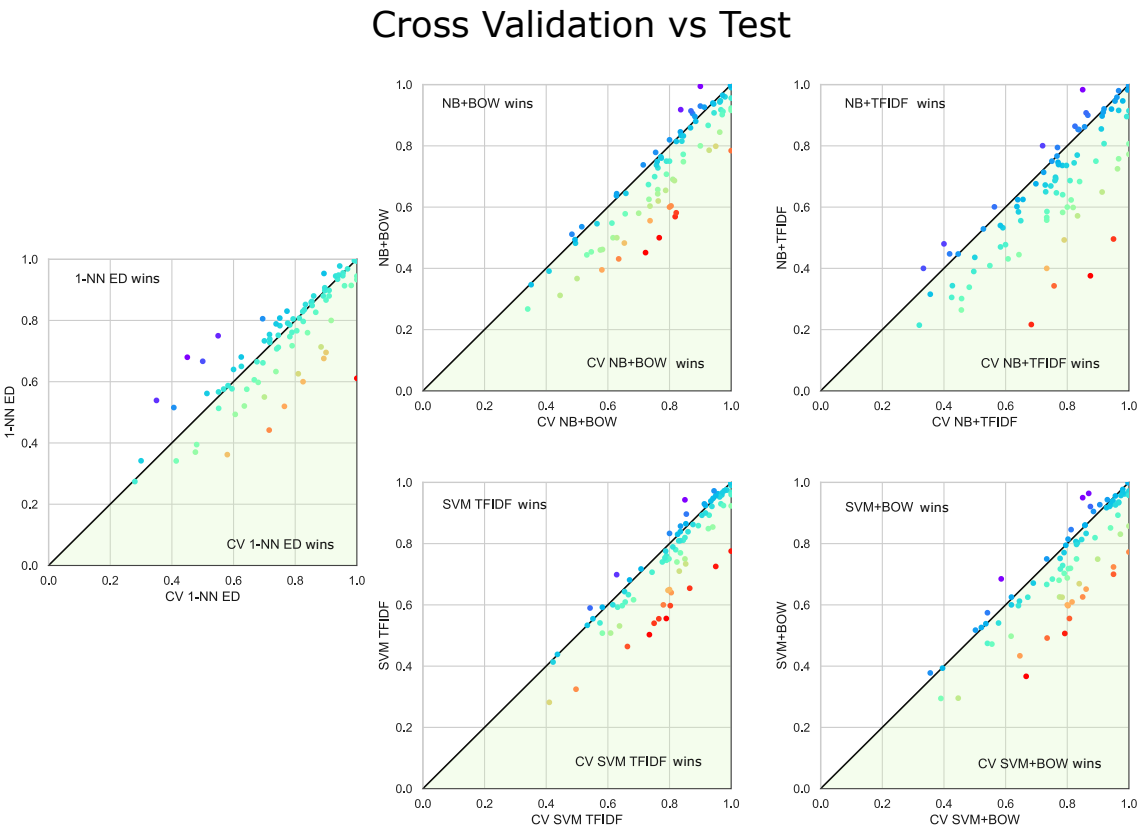


Figure B.6: Maps that compare the accuracy of each method during the cross-validation step and test step.

Table B.3: Parameter configuration and experimental results of each signal in Dataset 6 (see Section A.1.5). W_{size}: window size; K: kernel size in ratio of the window size; O: overlap ratio of the window size. T: amplitude threshold for event detection; Novelty: our proposed method with the novelty function; WL: window-based segmentation; BS: binary segmentation. The tolerance used to calculate the F1-score ($F1$) equals the window size for novelty function computation.

	Novelty					WL		BS
Signal	W_{size}	$K()$	$O()$	$T()$	$F1$	W_{size}	$F1$	$F1$
1	2000	0.62	0.95	0.19	0.90	2000	0.72	0.29
2	2000	1.25	0.95	0.28	0.86	1500	0.75	0.39
3	1500	1.27	0.95	0.10	1.00	2000	0.95	0.78
4	1500	0.63	0.95	0.14	1.00	2000	0.89	0.49
5	2000	1.25	0.95	0.28	0.92	1500	0.85	0.20
6	1500	0.63	0.95	0.28	0.88	1500	0.74	0.29
7	2000	1.25	0.95	0.19	0.95	2000	0.95	0.49
8	1500	1.27	0.95	0.10	0.93	2000	0.85	0.63
9	2000	0.62	0.95	0.28	1.00	2000	0.95	0.44
10	2000	1.25	0.95	0.28	0.97	1500	0.90	0.54
11	1000	1.25	0.95	0.10	1.00	2000	0.96	0.82
12	2000	1.25	0.95	0.14	0.83	2000	0.85	0.34
13	1000	0.62	0.95	0.37	1.00	2000	0.75	0.78
14	1000	1.25	0.95	0.32	0.95	1500	0.65	0.29
15	2000	0.62	0.95	0.14	0.93	1500	0.85	0.54
16	2000	0.62	0.95	0.28	0.90	2000	0.81	0.39
17	1500	2.53	0.95	0.10	0.98	2000	0.83	0.29
18	2000	0.62	0.95	0.50	0.97	1500	0.90	0.49
19	2000	0.62	0.95	0.23	0.93	1500	0.75	0.44
20	1500	0.63	0.95	0.10	0.95	2000	0.95	0.59
21	2000	1.25	0.95	0.19	0.95	2000	0.95	0.44
22	2000	1.25	0.95	0.19	0.84	2000	0.90	0.29
23	1500	1.27	0.95	0.28	0.90	1500	0.85	0.54
24	2000	0.62	0.95	0.19	0.86	2000	0.76	0.39
25	1500	1.27	0.95	0.37	0.95	2000	0.85	0.68
26	1500	0.63	0.95	0.19	0.98	2000	0.90	0.54
27	1500	1.27	0.95	0.19	0.83	1500	0.75	0.59
28	1500	1.27	0.95	0.37	0.90	1500	0.80	0.29
29	1500	1.27	0.95	0.23	1.00	2000	0.76	0.34
30	2000	2.5	0.95	0.10	0.92	2000	0.83	0.20
31	1000	1.25	0.95	0.19	1.00	1500	0.90	0.73
32	1500	1.27	0.95	0.41	0.89	2000	0.77	0.29
33	2000	0.62	0.95	0.23	0.83	1500	0.80	0.63
34	2000	0.62	0.95	0.28	0.93	1000	0.70	0.34
35	1500	0.63	0.95	0.14	0.95	2000	0.84	0.83
36	1000	2.5	0.95	0.14	1.00	1500	0.80	0.29

Table B.4: Parameter configuration and experimental results of each signal in Dataset 7 (see Section A.1.6). W_{size}: window size; K: kernel size in ratio of the window size; O: overlap ratio of the window size. T: amplitude threshold for event detection; Novelty: our proposed method with the novelty function; WL: window-based segmentation; BS: binary segmentation. The tolerance used to calculate the F1-score ($F1$) equals the window size for novelty function computation.

	Novelty					WL		BS
Signal	W_{size}	K	O	T	$F1$	W_{size}	$F1$	$F1$
1	1500	6.33	0.95	0.14	1.00	5000	0.93	0.97
2	5000	6.25	0.95	0.19	0.97	5000	0.79	0.97
3	5000	12.5	0.95	0.23	0.80	5000	0.57	0.83
4	3500	12.57	0.95	0.23	0.50	3500	0.21	0.48
5	5000	12.5	0.95	0.23	0.42	3500	0.14	0.34
6	1500	6.33	0.95	0.10	1.00	2500	0.93	0.90
7	2500	2.52	0.95	0.28	1.00	3500	0.86	0.90
8	2500	6.3	0.95	0.19	1.00	3500	0.64	0.90
9	2500	6.3	0.95	0.28	0.87	2500	0.29	0.55
10	5000	6.25	0.95	0.28	0.63	5000	0.21	0.34
11	5000	6.25	0.95	0.37	0.36	2500	0.07	0.34
12	2500	2.52	0.95	0.19	1.00	3500	0.93	0.90

Table B.5: Parameter configuration and experimental results of each signal in Dataset 8 (see Section A.1.7). W_{size}: window size; K: kernel size in ratio of the window size; O: overlap ratio of the window size. T: amplitude threshold for event detection; Novelty: our proposed method with the novelty function; WL: window-based segmentation; BS: binary segmentation. The tolerance used to calculate the F1-score ($F1$) equals the window size for novelty function computation.

	Novelty					WL		BS
Signal	W_{size}	K	O	T	$F1$	W_{size}	$F1$	$F1$
1	400	3.74	0.95	0.55	1.00	2000	0.67	0.40
2	400	2.49	0.95	0.68	0.80	400	1.00	0.40
3	150	4	0.95	0.82	0.80	150	0.00	0.00
4	150	2.66	0.95	0.32	1.00	2000	0.67	0.80
5	300	2.66	0.95	0.32	1.00	300	1.00	0.80
6	300	2.66	0.95	0.59	1.00	1000	0.67	0.80
7	150	2.66	0.95	0.64	1.00	2000	0.67	0.80
8	150	2.66	0.95	0.28	1.00	150	0.50	0.40
9	150	2.66	0.95	0.55	1.00	150	0.00	0.80

Table B.7: Parameter configuration and novelty function-based experimental results of each signal in Dataset 9 (*ATCPD*, see Section A.1.9). W_{size}: window size; K: kernel size in ratio of the window size; T: amplitude threshold for event detection; The tolerance corresponded to five samples, specified by the benchmark specifications [26].

Signal	$F1$	W_{size}	K	T
apple	0.95	10	50	0.65
bank	0.67	100	100	0.90
bee_waggle_6	0.66	100	250	0.80
bitcoin	0.69	10	65	0.15
brent_spot	0.86	10	30	0.75
businv	0.93	20	15	0.70
centralia	0.98	6	2	0.70
children_per_woman	0.88	10	10	0.70
co2_canada	0.85	10	20	0.35
construction	0.93	20	50	0.70
debt_ireland	0.97	6	2	0.70
gdp_argentina	0.97	20	10	0.70
gdp_croatia	1.00	20	10	0.70
gdp_iran	0.92	10	10	0.70
gdp_japan	1.00	10	10	0.70
global_co2	0.62	100	50	0.70
homeruns	0.93	15	25	0.50
iceland_tourism	0.65	150	150	0.99
jfk_passengers	0.98	20	30	0.80
lga_passengers	0.89	20	30	0.80
measles	0.17	20	30	0.80
nile	1.00	20	30	0.80
occupancy	0.95	10	20	0.40
ozone	0.86	6	4	0.60
quality_control_1	1.00	6	20	0.60
quality_control_2	1.00	6	20	0.60
quality_control_3	1.00	20	30	0.80
quality_control_4	0.97	10	50	0.75
rail_lines	0.91	6	2	0.75
ratner_stock	0.93	6	50	0.75
robocalls	0.98	6	10	0.50
scanline_126007	0.89	6	10	0.30
scanline_42049	0.98	6	20	0.50
seatbelts	0.66	6	20	0.30
shanghai_license	0.98	20	10	0.80
unemployment_nl	0.82	4	6	0.27
usd_isk	0.91	6	25	0.30
us_population	0.93	4	6	0.27
well_log	0.81	10	15	0.40

Table B.8: Results of type event 1 (work period transition), discriminated per time serie samples. Measurements of P, R, F1 of each according sample from Dataset 14 (see Section A.1.13). Signals without changes in activity were not considered (N.A).

Signal	P	R	F1
Operator 1 Workstation 1	0,78	1	0,88
Operator 1 Workstation 2	1	1	1
Operator 2 Workstation 1&2	0,86	1	0,92
Operator 3 Workstation 1	0,80	1	0,89
Operator 3 Workstation 2	N.A	N.A	N.A
Operator 4 Workstation 1	1	1	1
Operator 4 Workstation 2	N.A	N.A	N.A
Operator 5 Workstation 1	1	1	1
Operator 5 Workstation 2	1	1	1
Operator 6 Workstation 1	N.A	N.A	N.A
Operator 6 Workstation 2	N.A	N.A	N.A
Average	0.92	1	0.96

Table B.9: Detected cycles results of the detection of cycling events, on Dataset 14 (see Section A.1.13).

Signal	Detected Cycles	Duration Error
Operator 1 Workstation 1	11/11	3.26s (3.04%)
Operator 1 Workstation 2	14/15	8.09s (7.55%)
Operator 2 Workstation 1	14/14	6.45s (6.40%)
Operator 2 Workstation 2	11/11	11.2s (11.39%)
Operator 3 Workstation 1	16/16	12.35s (11.79%)
Operator 3 Workstation 2	13/13	11.41s (10.68%)
Operator 4 Workstation 1	14/14	1.05s (0.4%)
Operator 4 Workstation 2	11/11	3.42s (3.32%)
Operator 5 Workstation 1	12/12	2.83s (2.85%)
Operator 5 Workstation 2	10/11	3.47s (3.45%)
Operator 6 Workstation 1	14/15	3.79s (3.74%)
Operator 6 Workstation 2	14/15	5.79s (5.73%)
Total	154/157	N.A.

Table B.10: Overall results when applying HeaRTS and 1-NN ED on dataset 1. The first column indicates the name of the dataset, and the remaining columns are the name of the algorithms used. V-A: Cross-validation (CV) BoW+NB; A: BoW+NB; V-B: CV BoW+SVM; B: BoW+SVM; V-C: CV TF-idf+NB; C: TF-idf+NB; V-D: CV TF-idf+SVM; D: TF-idf+svm; V-E: CV 1-NN ED; E: 1-NN ED.

Dataset	V-A	A	V-B	B	V-C	C	V-D	D	V-E	E
ArrowHead	0.72	0.45	0.86	0.65	0.83	0.57	0.81	0.64	0.92	0.80
Beef	0.50	0.37	0.67	0.37	0.33	0.40	0.53	0.53	0.50	0.67
BeetleFly	0.80	0.60	0.85	0.95	0.80	0.60	0.80	0.65	0.55	0.75
BirdChicken	0.80	0.75	0.95	0.70	0.75	0.75	0.85	0.75	0.70	0.55
BME	1.00	1.00	1.00	1.00	0.97	0.98	1.00	1.00	0.87	0.83
Car	0.62	0.50	0.73	0.75	0.68	0.22	0.68	0.62	0.70	0.73
CBF	0.97	0.92	0.97	0.94	0.97	0.76	0.97	0.93	0.83	0.85
Chinatown	1.00	0.78	1.00	0.77	1.00	0.77	1.00	0.78	1.00	0.94
ChlorineConcentration	0.58	0.46	0.62	0.60	0.59	0.53	0.63	0.59	0.63	0.65
Coffee	0.93	0.79	0.96	0.89	0.89	0.75	0.93	0.89	1.00	1.00
Computers	0.77	0.76	0.80	0.69	0.78	0.74	0.82	0.74	0.57	0.58
CricketX	0.49	0.48	0.54	0.54	0.48	0.34	0.57	0.54	0.59	0.58
CricketY	0.55	0.45	0.54	0.47	0.50	0.44	0.58	0.51	0.55	0.57
CricketZ	0.52	0.54	0.54	0.57	0.46	0.26	0.58	0.59	0.58	0.59
Crop	0.49	0.49	0.52	0.53	0.45	0.45	0.55	0.55	0.72	0.73
DiatomSizeReduction	0.88	0.91	0.94	0.92	0.88	0.38	0.94	0.92	0.94	0.93
DistalPhalanxOutlineAgeGroup	0.85	0.75	0.85	0.63	0.84	0.68	0.85	0.73	0.81	0.63
DistalPhalanxOutlineCorrect	0.82	0.57	0.79	0.70	0.79	0.49	0.83	0.71	0.79	0.72
DistalPhalanxTW	0.79	0.65	0.78	0.63	0.79	0.58	0.80	0.65	0.74	0.63
Earthquakes	0.81	0.69	0.81	0.72	0.83	0.77	0.84	0.77	0.75	0.71
ECG200	0.79	0.75	0.83	0.75	0.73	0.65	0.79	0.76	0.86	0.88
ECGFiveDays	0.87	0.91	0.87	0.96	0.91	0.65	0.96	0.96	0.83	0.80
EOGHorizontalSignal	0.64	0.43	0.73	0.49	0.64	0.44	0.73	0.50	0.72	0.44
EOGVerticalSignal	0.58	0.40	0.65	0.43	0.61	0.43	0.66	0.46	0.58	0.36
EthanolLevel	0.53	0.44	0.56	0.47	0.58	0.47	0.61	0.51	0.28	0.27
FaceAll	0.75	0.70	0.80	0.72	0.73	0.67	0.80	0.75	0.88	0.71
Fish	0.84	0.85	0.81	0.85	0.76	0.34	0.84	0.86	0.75	0.78
FordA	0.71	0.74	0.80	0.79	0.77	0.75	0.83	0.83	0.68	0.67
FordB	0.82	0.58	0.84	0.67	0.82	0.60	0.87	0.65	0.67	0.61
FreezerRegularTrain	0.97	0.91	0.98	0.98	0.97	0.92	0.99	0.97	0.79	0.80
FreezerSmallTrain	1.00	0.92	1.00	0.97	1.00	0.91	1.00	0.97	0.89	0.68
GunPointAgeSpan	0.97	0.90	0.99	0.96	0.99	0.90	0.97	0.96	0.97	0.97
GunPointMaleVersusFemale	0.96	0.94	0.96	0.96	0.95	0.95	0.97	0.97	1.00	0.99
GunPointOldVersusYoung	0.94	0.94	0.95	0.94	0.92	0.92	0.96	0.95	1.00	1.00
GunPoint	0.98	0.96	0.98	0.97	0.98	0.95	1.00	0.96	0.96	0.91
Ham	0.76	0.66	0.82	0.61	0.70	0.68	0.78	0.60	0.83	0.60
HandOutlines	0.82	0.69	0.86	0.86	0.81	0.69	0.85	0.86	0.85	0.86
Haptics	0.45	0.31	0.45	0.30	0.43	0.31	0.50	0.32	0.48	0.37
Herring	0.66	0.58	0.67	0.63	0.66	0.63	0.66	0.61	0.41	0.52
HouseTwenty	1.00	0.92	1.00	0.96	1.00	0.81	0.98	0.92	0.63	0.68
InlineSkate	0.34	0.27	0.39	0.29	0.32	0.21	0.41	0.28	0.30	0.34
InsectEPGRegularTrain	1.00	0.96	1.00	0.97	0.92	0.81	0.98	0.98	1.00	1.00
InsectEPGSmallTrain	0.88	0.90	0.94	0.93	0.76	0.65	0.94	0.95	1.00	1.00
InsectWingbeatSound	0.41	0.39	0.40	0.39	0.43	0.34	0.44	0.44	0.51	0.56
ItalyPowerDemand	0.93	0.87	0.94	0.85	0.91	0.86	0.94	0.85	0.96	0.96

Continues from Table B.10

Dataset	V-A	A	V-B	B	V-C	C	V-D	D	V-E	E
LargeKitchenAppliances	0.76	0.73	0.90	0.75	0.80	0.74	0.89	0.86	0.61	0.49
Lightning2	0.80	0.82	0.83	0.80	0.73	0.56	0.85	0.82	0.72	0.75
Lightning7	0.63	0.64	0.59	0.68	0.46	0.30	0.63	0.70	0.64	0.58
Mallat	0.84	0.92	0.93	0.92	0.56	0.60	0.85	0.90	0.98	0.91
Meat	0.90	0.80	0.97	0.93	0.87	0.90	0.92	0.90	1.00	0.93
MelbournePedestrian	0.61	0.55	0.64	0.61	0.58	0.54	0.65	0.64	0.93	0.95
MiddlePhalanxOutlineAgeGroup	0.81	0.60	0.80	0.60	0.81	0.62	0.80	0.60	0.77	0.52
MiddlePhalanxOutlineCorrect	0.73	0.63	0.77	0.68	0.76	0.70	0.79	0.75	0.81	0.77
MiddlePhalanxTW	0.63	0.50	0.64	0.60	0.64	0.58	0.65	0.60	0.55	0.51
MixedShapesSmallTrain	0.84	0.82	0.89	0.82	0.72	0.80	0.86	0.84	0.83	0.84
MixedShapes	0.87	0.86	0.95	0.94	0.84	0.85	0.93	0.94	0.90	0.90
NonInvasiveFetalECGThorax1	0.48	0.51	0.50	0.52	0.42	0.45	0.54	0.59	0.82	0.83
NonInvasiveFetalECGThorax2	0.63	0.64	0.62	0.63	0.50	0.39	0.67	0.68	0.89	0.88
OliveOil	0.77	0.50	0.73	0.67	0.73	0.40	0.67	0.63	0.90	0.87
OSULeaf	0.89	0.88	0.89	0.90	0.79	0.64	0.91	0.89	0.64	0.52
PhalangesOutlinesCorrect	0.76	0.64	0.75	0.68	0.75	0.67	0.78	0.74	0.79	0.76
PowerCons	0.84	0.77	0.86	0.83	0.82	0.74	0.84	0.81	0.94	0.98
ProximalPhalanxOutlineAgeGroup	0.84	0.83	0.80	0.81	0.84	0.85	0.84	0.84	0.78	0.79
ProximalPhalanxOutlineCorrect	0.73	0.67	0.79	0.73	0.77	0.74	0.82	0.78	0.81	0.81
ProximalPhalanxTW	0.79	0.71	0.78	0.72	0.76	0.69	0.79	0.78	0.74	0.71
RefrigerationDevices	0.65	0.48	0.79	0.51	0.60	0.48	0.77	0.55	0.48	0.39
Rock	0.70	0.58	0.80	0.60	0.40	0.48	0.75	0.54	0.60	0.64
SemgHandGenderCh2	0.76	0.62	0.78	0.63	0.76	0.69	0.74	0.71	0.91	0.90
SemgHandMovementCh2	0.35	0.35	0.36	0.38	0.36	0.32	0.42	0.41	0.68	0.60
SemgHandSubjectCh2	0.58	0.46	0.62	0.50	0.55	0.41	0.64	0.53	0.78	0.79
ShapeletSim	0.90	0.99	1.00	1.00	0.85	0.98	1.00	0.99	0.35	0.54
ShapesAll	0.74	0.60	0.78	0.75	0.64	0.60	0.79	0.75	0.75	0.75
SmallKitchenAppliances	0.75	0.78	0.79	0.77	0.77	0.79	0.81	0.79	0.41	0.34
SmoothSubspace	0.77	0.76	0.85	0.81	0.77	0.77	0.80	0.83	0.89	0.95
SonyAIBORobotSurface1	0.95	0.80	0.95	0.72	0.95	0.50	0.95	0.73	0.90	0.70
SonyAIBORobotSurface2	0.96	0.84	1.00	0.86	0.96	0.73	0.93	0.85	0.85	0.86
StarlightCurves	0.97	0.95	0.97	0.96	0.96	0.96	0.98	0.97	0.86	0.85
Strawberry	0.84	0.83	0.91	0.93	0.86	0.86	0.92	0.92	0.95	0.95
SwedishLeaf	0.82	0.81	0.86	0.86	0.73	0.59	0.88	0.86	0.74	0.79
SyntheticControl	0.94	0.94	0.93	0.94	0.94	0.92	0.95	0.96	0.91	0.88
ToeSegmentation1	0.90	0.93	0.88	0.92	0.83	0.86	0.85	0.94	0.45	0.68
ToeSegmentation2	0.94	0.91	0.97	0.83	0.86	0.91	1.00	0.92	0.75	0.81
Trace	1.00	0.99	1.00	1.00	1.00	0.99	1.00	0.99	0.84	0.76
TwoLeadECG	0.91	0.93	0.96	0.93	0.91	0.90	0.91	0.93	0.78	0.75
TwoPatterns	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.91
UMD	0.97	0.97	1.00	0.96	0.92	0.91	0.94	0.97	0.69	0.81
UWaveGestureLibraryAll	0.76	0.74	0.83	0.81	0.73	0.57	0.83	0.81	0.94	0.95
UWaveGestureLibraryX	0.66	0.64	0.69	0.67	0.64	0.62	0.71	0.72	0.72	0.74
UWaveGestureLibraryY	0.56	0.55	0.58	0.54	0.53	0.53	0.62	0.60	0.70	0.66
Wafer	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.99	1.00
Wine	0.74	0.56	0.81	0.56	0.65	0.56	0.79	0.56	1.00	0.61
Yoga	0.76	0.75	0.83	0.80	0.72	0.71	0.83	0.81	0.77	0.83
Average	0.78	0.71	0.81	0.75	0.75	0.66	0.81	0.76	0.76	0.74