Pedro Miguel Alves da Silva

Bachelor of Science in Biomedical Engineering

Clinical deterioration detection for continuous vital signs monitoring using wearable sensors

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Biomedical Engineering

Adviser: Prof. Dr. Hermie Hermens, Full Professor, University of Twente

Co-adviser: Prof. Dr. Carla Maria Quintão Pereira, Assistant Professor, NOVA School of Science and Technology, NOVA University of Lisbon

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

February, 2021

Clinical deterioration detection for continuous vital signs monitoring using wearable sensors

This document was created using the (pdf)LaTeX processor, based on the NOVAthesis template, developed at the Dep. Informática of FCT-NOVA by João M. Lourenço.

To my parents, my grandparents and my sister.

Acknowledgements

This work marks the end of a five years journey for me, which wouldn't have been the same without some people. In the end, the merit will be attributed to me, but the credits are theirs too.

In first place, thank you Professor Hermie Hermens for the opportunity to work in BSS and to experience life at UT. To Mathilde Hermans I would like to express my word of gratitude, for all your supervision and guidance throughout this thesis. Your feedback, help and constant motivation were extremely valuable to the success of this work. Your kindness, friendship and enthusiasm made this whole experience much more gratifying and pleasant. Furthermore, I thank all my BSS coworkers for the companionship and tea breaks we had together.

Then, I leave a word of appreciation to Professor Carla Quintão. Thank you for your availability and concern during this thesis. Also, I would like to seize this opportunity to express my tremendous admiration for your professionalism and for the dedication you put in everyday for the Biomedical Engineering students at FCT. You always looked out for our best interests, so thank you very much for that.

To all my friends, a huge thanks for all these years together. I couldn't have done it without you, for each and every one of you were a fundamental part of this journey. You kept my motivation and perseverance high, you were there for me when things were tough, you gave me the best memories and, more importantly, you made me happier and helped shape who I am today.

To Mini, my ugliest and bravest partner and wingman; To Bessa, my lifetime brother; To the remaining *Mosguerreiros*, Fitas, Fabinho, Félix and Sousa, thank you for all the vacations, stupidity, ugliness and laughter we went through together.

To Xico, my twin and soulmate; To Luís, my football sidekick and "logic-only" partner; To Johnny Afonso, my craziest brother, with whom I always have a good time; To Tufão, Igor, Benny, Daniel and the remaining *Javalis*, thank you for all the friendship, the tough nights, the wild trips and understanding when I disappeared for months to study.

To Afonso, my other pea in the life pod; To the remaining *P&P*, João, Romão and Saraiva, thank you for all the jokery, the deep talks, for making me less of a toaster and for an awesome high school.

To my godchildren, Sarah, my physics brain-killer and favorite psychology client, and Rosinha, my future rockstar with lead feet, thank you for making me proud everyday and

for showing me that with enough power of will you can do everything. To all my college friends, Garcia, Toni, Teresa, Gonçalo, Zagalo, André, Correia, Kika, Bernardo, David, Saren, Rita, Ana Maria, Sara Santos, Diogo and all the other incredible people I met on the best course, thank you for the best five years ever, it was an amazing journey thanks to you. To my Erasmus friends, thank you very much for being part of such a remarkable experience for me.

Finally, I would like to give the biggest word of gratitude to my family. My parents, Olga and Aníbal, you were my greatest supporters during these five years. You provided me with everything a son could ask for and were always patient with me, even in my moody four-tests weeks. You are the definition of a role model, hard work and love. Thank you for being the best parents in the world. To my sister, Nonô, you are the reason I push myself harder everyday. You make me want to be the best example a little sister can have. Without you and your support, I would never have achieved everything I did. Also, thank you for suffering with me when Benfica loses and for never complaining about the music while I study. The three of you are my safe place, mean the world to me and have provided me with 23 years of happiness. And for that, I will always be thankful. To my grandparents, Lurdes, Florinda, Fernando and Manuel, thank you for everything you did for me, for the delicious meals and for the unconditional support. To the rest of my family, my uncles and aunts, Sónia, Arlindo, Elsa, Ana and Jorge, my cousins, Rodrigo, Luís, Daniel, Sofia, Mariana, Juliana, Duarte and Diogo, my borrowed uncles and aunts, Antonieta, Quim, Cristina and Rui, thank you for all your love and support.

The good thing about science is that it's true whether or not you believe in it

- Neil deGrasse Tyson

Abstract

Surgical patients are at risk of experiencing clinical deterioration events, especially when transferred to general wards during the postoperative period of their hospital stay. Currently, such events are detected by combining Early Warning Scores (EWS) with manual and periodical vital signs measurements, performed by nurses every 4 to 6 hours. Hence, deterioration may remain unnoticed for hours, delaying patient treatment, which might lead to increased morbidity and mortality. Also, EWS are inadequate to predict events so physiologically complex.

So that early warning of deterioration could be provided, it was investigated the potential of warning systems that combine machine learning-based prediction models with continuous vital signs monitoring, provided by wearable sensors.

This dissertation presents the development of such a warning system, fully independent of manual measurements and based on a logistic regression prediction model with 85% sensitivity, 79% precision and 98% specificity. Additionally, a new personalized approach to handle missing data periods in vital signs and a novel variation of a RR-interval preprocessing technique were developed. The results obtained revealed a relevant improvement in the detection of deterioration events and a significant reduction in false alarms, when comparing the warning system with a commonly employed EWS (42% sensitivity, 14% precision and 90% specificity). It was also found that the developed system can assess patient's condition much more frequently and with timely deterioration detection, without even requiring nurses to interrupt their workflow. These findings support the idea that these warning systems are reliable, more practical, more appropriate and produce smarter alarms than current methods, making early deterioration detection possible, thus contributing for better patients outcomes. Nonetheless, the performance achieved may yet reveal insufficient for application in real clinical contexts. Therefore, further work is necessary to improve prediction performance to a greater extent and to confirm these systems reliability.

Keywords: Clinical deterioration, Continuous monitoring, Wearable sensors, Vital signs, Machine learning, Warning system.

Resumo

Pacientes cirúrgicos estão em risco de experienciar eventos de deterioração clínica, especialmente quando transferidos para alas gerais durante o período pós-operatório da sua estadia hospitalar. Atualmente, esses eventos são detetados através da combinação de Pontuações de Alerta Antecipado (PAA) com medições manuais de sinais vitais, realizadas por enfermeiros a cada 4 a 6 horas. Consequentemente, deterioração pode passar despercebida durante horas, atrasando o tratamento do paciente, podendo levar a morbidade e mortalidade aumentada. Para que alertas antecipados de deterioração possam ser fornecidos, foi investigado o potencial de sistemas de alerta que combinam modelos de predição baseados em aprendizagem automática, com monitorização contínua de sinais vitais, proporcionada por sensores vestíveis. Esta dissertação apresenta o desenvolvimento de tal sistema de alerta, totalmente independente de medições manuais e baseado num modelo de predição de regressão logística com 85% sensibilidade, 79% precisão e 98% especificidade. Além disso, uma abordagem nova e personalizada para lidar com períodos de ausência de dados nos sinais vitais e uma nova variação de uma técnica de pré-processamento do intervalo R-R foram desenvolvidas. Os resultados obtidos revelaram uma melhoria na deteção de eventos de deterioração e uma redução significativa de alarmes falsos, quando comparando o sistema de alerta com uma PAA regularmente usada (42% sensibilidade, 14% precisão e 90% especificidade). Também foi descoberto que o sistema desenvolvido pode avaliar a condição de um paciente mais frequentemente, sem necessitar que enfermeiros interropam a sua atividade. Estas descobertas suportam a ideia de que estes sistemas de alerta são fidedignos, mais práticos e produzem alarmes mais inteligentes que os métodos atuais, tornando possível a deteção antecipada de deterioração, contribuindo para melhores desfechos médicos dos pacientes. No entanto, o desempenho alcançado pode revelar-se ainda insuficiente para aplicação em contextos clínicos reais. Por esse motivo, estudos futuros são necessários para melhorar ainda mais o desempenho na predição e para confirmar a confiabilidade destes sistemas.

Palavras-chave: Deterioração clínica, Monitorização contínua, Sensores vestíveis, Sinais vitais, Aprendizagem automática, Sistema de alerta.

Contents

List of Figures

List of Tables

Acronyms and abbreviations

AE Adverse events

ASA American Society of Anesthesiologists class

AUC Area Under the receiver operating characteristic Curve

BP Blood Pressure

BT Boosted Trees

BTemp Body Temperature

ECG Electrocardiogram

EWS Early Warning Scores

HR Heart Rate

ICU Intensive Care Units

LinInt Linear Interpolation dataset

LR Logistic Regression

MEWS Modified Early Warning Score

ML Machine Learning

NApp1 New approach version 1 dataset

NApp2 New approach version 2 dataset

PPG Photoplethysmogram

QRSa QRS complex amplitude

RR Respiration Rate

RRI RR interval

SAE Serious adverse events

SpO_2 (Peripheral) Oxygen Saturation

UGI Upper gastrointestinal

Chapter 1

Introduction

1.1 Problem and context

It is known that there is a growing trend to transfer surgical patients[1] from Intensive Care Units (ICU) to general wards earlier [1]. Additionally, there are indications that these patients are progressively becoming more frail [2], [3] and it was recently reported that almost 50% of all Adverse events (AE) occur in the general ward [4]. These facts are all contributing for the aggravation of the problem in hands, which originates from the occurrence of clinical deterioration events, in surgical patients, due to the limitations presented by current monitoring systems employed in general wards. Besides the poor monitoring conditions in these wards, the nurse-to-patient ratio is also much worse [5]. In fact, that ratio can go from 1:1 in ICU [6] to 1:4 or even 1:10 in general wards [7], [8]. Also, the National Institute for Health and Care Excellence recommends vital signs measurements every 12 hours as a minimum [9]. In practice, this is usually performed every 4 to 6 hours [5], [10], which still leaves large time gaps without monitoring, hence leading to the first signs of complications to be unnoticed for hours [5], [11]. Besides this low frequency of vital signs measurements, incomplete vital signs recordings, errors calculating EWS (EWS explanation in chapter 3) and inconsistent calls for rapid response teams were suggested as causes for the delay in recognizing deteriorating patients and for the missing of such events [12]–[14].

Most complications that come from surgical interventions are due to postoperative developments rather than pre- or intra-operative ones. Also, it was reported that around 25% of surgical patients suffer with postoperative deterioration events [5] and that 73% of those who died after surgery, were never transferred into a higher level of care [15]. This continues to happen because currently, patient condition is assessed by combining

[1] surgical patients are patients that are submitted to surgical interventions

intermittent routine nurse controls with EWS, which is a strategy that presents some disadvantages. Firstly, these scores cannot fully cope with the complexity of these patients' physiology [5]. Secondly, this process relies on the nurses' observations, which are subjective, error prone and time-consuming. Finally, its periodical nature might miss important events occurring between measurements [5], [16]. Even by allying EWS with rapid response teams, efficiency in preventing clinical deterioration has failed to increase [17].

Poor clinical monitoring is, thus, a major issue when clinical deterioration prevention is being discussed. In fact, it has been shown to be the main reason for preventable deaths to still take place in acute hospitals [18]. In their studies, Hogan et al. [19] reported that 31.3% of preventable deaths were due to improper clinical monitoring, while Taenzer et al. [20] found out that 29% of respiratory depression events were also related to inadequate monitoring. Also, Sun et al. [21] reported that 90% of hypoxemic episodes were missed using periodical monitoring. Other studies suggest that where there's more margin of improvement is not in the treatment of the deteriorating patient but in its early identification [22]. All of these studies support the idea that the current state of clinical deterioration detection strategies causes delays in the identification of complications. This delay can lead to increased morbidity, AE, Serious adverse events (SAE), undesired outcomes, enhanced hospitalization costs and length of stay, unplanned readmission to ICU and higher mortality rates [23]–[25]. Considering that the underlying problems that lead to these outcomes manifest hours prior to a complication and are known to be preceded by abnormal changes in vital signs [20], [25]–[28], adequate monitoring and early detection can be the key to avoid the evolution of these health detrimental events.

The solution to achieve earlier detection might be continuous vital signs monitoring. In fact, its implementation in general wards has already been recommended [5]. However, this approach hasn't yet proven to have or not an effect on reducing the number of undesired clinical outcomes and requires further research [11]. While some studies obtained promising results others didn't [2], [11], [20], [27]. The reason for this is that most studies focus their attention on the entire action chain[2], whilst, for this matter, they should only be interested in the afferent limb, because the response protocol affects clinical outcome too [29]. Also, others [2], [30], [31] only used SAE as metric to conclude if continuous monitoring is valuable, which never gets to happen for some patients that can still take advantage of this approach. So, for a more valid and complete evaluation of continuous monitoring utility, (1) AE should be considered as well; (2) the evaluation should only concern whether this method detects abnormalities earlier than nurses intermittent observations, i.e., contemplating the afferent limb only.

Some predictive models, that are based on continuous vital signs measurements and are often more complex than EWS, have already proven to enhance the capability of identifying deterioration and to be feasible [32], [33] in general wards settings. So, the

[2]the action chain can be divided into two components: the afferent limb, that involves monitoring to detect deterioration, and the efferent limb, that involves responding to this deterioration [5]

possible contribution of these decision support models for predicting deterioration events has already started to be investigated, remaining yet unclear the role that demographic and contextual information, like the number of comorbidities or age, can have. In chapter 3, a review of these models is performed in more detail.

These predictive models can be particularly useful for large scale applications if combined with low cost wearable sensors. Recently, this technology has been introduced in order to aid the earlier assessment of clinical deterioration, by allowing the detection of abnormal vital signs trends and by complementing nurses' observations, which hastens intervention and minimizes impairments. The recent emergence of these devices materialized as a result of advancements in information and communication technologies, advancements in biomedical signal processing techniques and due to the need for improved healthcare delivery [10], [25]. Besides allowing patient monitoring to be made continuously, these sensors also assure freedom of movement to the patients, which promotes acceptability and enables quicker recoveries [10]. However, there's still concerns about the existing alarm systems, used in conjunction with these devices, having too many false alarms [5], [11], which eventually leads to alarm fatigue[3] in the nursing staff. This phenomenon makes the development of new decision support models imperative. Gross et al. [34] concluded that only by adjusting alarm limits to the population, a significant reduction in false alarms can be reached. Other issues that still need to be addressed regarding this technology are its reliability, quality of measurements, power management, security and patient confidentiality [10].

By combining wearable devices with continuous monitoring and an adequate decision support model, it is expected to create systems that can automate patient monitoring in general wards, where less advanced equipment is present and where the monitoring conditions are worse, improving workflow and patient's outcomes. These systems should fulfill the following requirements: minimal patient burden caused by the sensors, the system must support caregivers in interpreting the large amounts of data provided by continuous monitoring, automatic identification of anomalies or undesired trends in the continuous data, alerts in case of patient deterioration, integration with the Electronic Medical Record and high performance in the deterioration detection task.

Study populations

Every patient can deteriorate. Hence, every patient can benefit from these monitoring systems. However, depending on the underlying reason that led to a patient's hospitalization, the risk of developing complications varies. Surgical patients, and particularly, patients undergoing Upper gastrointestinal (UGI) surgery and geriatric patients undergoing hip fracture surgery are amongst the groups with higher rates of deterioration

[3] alarm fatigue can be defined as the process of clinical staff becoming desensitized to, and ultimately ignoring, alerts from monitoring systems [10]

development during ward stay. In the first case, this is associated with a high inflammatory response both during and after surgery [35]. In fact, it is expected that between 20% and 80% of these patients develop postoperative complications such as pneumonia and other pulmonary conditions, anastomotic leaks and cardiac complications [36]. Most UGI surgeries are gastroesophageal cancer resections and a 2016 study [37] on the subject reported a postoperative mortality from 1% to 7%. Two other studies [38], [39] reported failure-to-rescue[4] rates of 7% to 24%. It's worth to mention that gastroesophageal cancer is a very serious disease that affected around 1.5 million people worldwide in 2011 alone [35]. In the second population, the high rate of complications is associated with the patients' age and medical condition. In a 2017 study [40], it was reported a 50% rate of pre- or postoperative complications development, such as immobility and mortality, in these elderly patients. This condition has an annual incidence of between 0.25% and 2.5% in people older than 60 years, in the United States of America and Europe [41].

1.2 Goals

This dissertation project focused on the afferent limb, i.e., on the part of the action chain that aims to detect signs of clinical deterioration. The main research question associated with this study is:

Is it possible to develop a warning system, that performs better than current ones, in the early detection of deterioration in surgical patients that are monitored continuously using wearable vital signs sensors in general wards?

By doing so, it's expected that deterioration is perceived earlier and, consequently, further damage to the patients can be prevented, both enhancing patient's health and reducing hospital costs and length of stay. Also, the number of false alarms is foreseen to diminish, which will lessen alarm fatigue in nurses.

To achieve this, several more specific goals were established:

1. **Review recent studies regarding the development of clinical deterioration prediction models** - a thorough literature review on previously described methods for the development of decision support systems was crucial to identify the limitations that still had to be tackled. In addition, this provides a more complete comparison between previous studies and this project's results and design.

2. **Identification and improvement of existing preprocessing techniques** - commonly implemented preprocessing techniques for the different types of data continuously recorded, and used in this thesis (see subsection 4.2.2), had to be identified. In order to improve the performance in the prediction task and, consequently, the quality of

[4] mortality in patients that had complications

the final warning system, the implementation of innovative solutions for some of the preprocessing steps was also a goal of this project.

3. **Identification of predictors/features that indicate the presence of complications** – given that an extensive set of features can be found in the literature, it was imperative to identify which ones can actually provide insight about patterns of deterioration. An additional objective was to assess if features based on correlations between vital signs, and demographic and contextual features could also be predictors of deterioration.

4. **Development of a decision support model** – implementation of a prediction model that supports data interpretation and automatically warns clinicians in case of complications, based on Machine Learning (ML) algorithms that make use of the features previously identified.

This study used a dataset with unique context information, that had not yet been studied, and sought to tackle all the limitations present in previous studies (see section 3.2). Other unique characteristics of this work were: (1) the topics chosen to be summarized in the literature review (see appendix A); (2) the analysis of which sensors were unreliable and the development of an additional model that didn't require the use of variables measured through those sensors; (3) the novel preprocessing techniques developed; (4) the measuring of the time it takes for each patient assessment to be completed by the warning system, in order to report an estimation of a realistic usage frequency in a real clinical context.

1.3 Thesis outline

This thesis is composed by 7 chapters and 3 appendices. In the present chapter, the associated problem was contextualized and the goals of the project were delineated. In chapter 2, relevant theoretical concepts for the understanding of the remaining work are described. In chapter 3, the current methods employed for clinical deterioration detection, and their respective limitations, are presented. Additionally, a thorough review is performed on novel strategies implemented in recent studies. Chapter 4 specifies the sensors used to acquire the data and provides a dataset description. In chapter 5, the methodology employed for preprocessing the continuously acquired data is presented, as well as related results. Chapter 6 describes the development process of the warning system and respective assembly. This includes the methodology and the results associated with the decision support model development. The study's limitations and suggestions for future work in the field are also provided. In chapter 7, the final remarks and main achievements are emphasized. Appendix A is a summary of the work reviewed in chapter 3, in a table format. Appendix B describes the features extracted for the development of

decision support models. Finally, appendix C is a set of additional results, not included in the main text, mostly because of their dimensions.

Chapter 2

Theoretical concepts

In this chapter, the most important concepts mentioned throughout the thesis are detailed. First, clinical deterioration is defined and related events of interest are specified. Then, two types of monitoring are described, vital signs and ECG monitoring, which are accompanied by an explanation of the variables that were extracted from each of them. Lastly, the Machine Learning (ML) field is approached, with focus on prediction models and the algorithms used to explore their development, and on important metrics for this kind of applications.

2.1 Clinical deterioration

There's a lack of accordance when trying to define what's clinical deterioration. This fact has implications in its recognition and respective response quality [42], [43].

According to Jones et al. [43], clinical deterioration is a term that has changed over time. Early definitions focused on the end result, i.e., used consequences as sepsis or cardiac arrest to define deterioration. Current definitions state that the term is related with abnormalities in vital signs and other clinical parameters, and that those are seen as an assistive tool for clinicians to prevent subsequent risk.

Padilla et al. [42] defined clinical deterioration as "a dynamic state experienced by a patient compromising hemodynamic stability, marked by physiological decompensation accompanied by subjective or objective findings". It was also concluded that its identification is a crucial factor to inpatient mortality, that it can occur in any stage of a patient's hospitalization and that contextual factors might play an important role.

Whatever definition is adopted, three underlying elements are always present: appearance of anomalies in the patient's vital signs; likelihood of adverse outcomes occurrence; consequences include mortality, transfer to a higher level of care and prolonged hospital

stay.

Usually, the health related events that cause clinical deterioration are denominated Adverse events (AE), which includes Serious adverse events (SAE).

An AE can be defined as an unexpected medical complication suffered from a patient during a study/hospital stay [44]. These don't necessarily have to be linked with the study or with the primary problem. For this project, AE of interest were pre- or postoperative complications with a Clavien Dindo class of II or higher (Table 2.1), diagnosed according to standard guidelines.

Table 2.1: Classification of surgical complications [45].

Grade	Definition
Grade I	Any deviation from the normal postoperative course without the need for pharmacological treatment or surgical, endoscopic, and radiological interventions
	Allowed therapeutic regimens are: drugs as antiemetics, antipyretics, analgetics, diuretics, electrolytes, and physiotherapy. This grade also includes wound infections opened at the bedside
Grade II	Requiring pharmacological treatment with drugs other than such allowed for grade I complications
	Blood transfusions and total parenteral nutrition are also included
Grade III	Requiring surgical, endoscopic or radiological intervention
Grade IIIa	Intervention not under general anesthesia
Grade IIIb	Intervention under general anesthesia
Grade IV	Life-threatening complication (including CNS complications)* requiring IC/ICU management
Grade IVa	Single organ dysfunction (including dialysis)
Grade IVb	Multiorgan dysfunction
Grade V	Death of a patient
Suffix "d"	If the patient suffers from a complication at the time of discharge (see examples in Table 2), the suffix "d" (for "disability") is added to the respective grade of complication. This label indicates the need for a follow-up to fully evaluate the complication.

*Brain hemorrhage, ischemic stroke, subarrachnoidal bleeding, but excluding transient ischemic attacks.
CNS, central nervous system; IC, intermediate care; ICU, intensive care unit.

Some commonly mentioned SAE are myocardial infarction, renal failure, cardiac arrest, severe sepsis, unexpected death or events that lead to unplanned ICU admission or emergency surgery. FDA [46] has defined SAE as the ones that: result in death; are life threatening (at the time of the event); require (prolonged) hospitalization; result in permanent or significant disability or incapacity; result in a congenital anomaly or birth defect; required intervention to prevent permanent impairment or damage; any other important medical event that did not result in any of the outcomes listed above, only due to medical or surgical intervention.

2.2 Vital signs monitoring

This expression refers to the intermittent or continuous observation of a patient's vital signs in order to assure its safety and guide therapeutic procedures [4]. Vital signs are medical signs that provide insight regarding the body's vital functions condition. As said before, clinical deterioration is associated with the presence of abnormalities in a patient's vital signs. Thereby, appropriate vital signs monitoring is crucial to achieve a positive patient outcome. Usually, Blood Pressure (BP), Body Temperature (BTemp),

Heart Rate (HR), Respiration Rate (RR) and (Peripheral) Oxygen Saturation (SpO_2) are referred to as the five vital signs [2], [28] and can be monitored with automated equipment or manually by intermittent nurse checks. Despite their importance, little is known about what the best regimes and frequencies are to measure them [2], [28], [47]. Nonetheless, the conventional measuring frequency, currently in practice in general wards, is once every 4 to 6 hours [5], [10]. To surpass this and other issues regarding intermittent monitoring (see chapter 1), wearable devices, that can continuously monitor patients for long periods of time, are gaining momentum, which might lead to the use of these devices becoming the standard choice for vital signs monitoring.

2.2.1 Blood pressure

BP is the pressure that circulating blood exerts on the walls of arteries [48]. Its measurement provides two values: systolic BP and diastolic BP. Their normal values vary with age and gender. Deviant values can cause atherosclerosis, strokes or heart attacks, and can be caused by severe infections, heart problems, unhealthy lifestyle, amongst others. It can be measured through manual sphygmomanometers or with digital devices placed in the upper arm, wrist or finger.

2.2.2 Body temperature

BTemp is the average temperature of the human body. Its typical value is around 37 °C but can vary throughout the circadian cycle and among individuals. Irregularities in this parameter are defined as hyperthermia (higher values) and hypothermia (lower values) [49] and these can have severe consequences like heat rashes, heat cramps, heat exhaustion, heat stroke, neurological dysfunctions, myocardial ischemia and death. These deviances can be induced by thermal stress (environment temperature or physical exercise), metabolic disorders, infections or drugs [50]. BTemp can be measured with thermometers in different body regions: ear, mouth, rectum and armpit.

2.2.3 Heart rate

HR is the number of heartbeats (ventricles' contractions) per time unit. It's mostly reported in bpm (beats per minute). Its normal range of values for an adult goes from 60 bpm to 100 bpm, although it varies with activity levels, age, gender or even fitness levels. If an individual's HR is higher (tachycardia) or lower (bradycardia) than expected, some of the causes might be related to heart and heart's conduction pathways disorders, stress or medicinal drugs. These conditions' consequences may be sudden cardiac arrest, heart failure, syncope or strokes. In wards, HR is usually measured manually by nurses through pulse counts or resorting to the ECG, Photoplethysmogram (PPG) or accelerometry signals [51].

2.2.4 (Peripheral) Oxygen saturation

SpO_2 is a measure of the proportion of total hemoglobin that is oxyhemoglobin. Its healthy values range from 95% to 100%, where values below that might be considered as signs of hypoxemia (there's no standard threshold defined though) [52]. Hypoxemia's major causes are anemia and heart and lung conditions. It might result in hypoxia and tissue damage. In wards, SpO_2 can be continuously measured by pulse oximetry, usually through finger-based probes.

2.2.5 Respiration rate

RR is defined as the number of breaths per time unit, i.e., the number of inspirations or expirations, generally, per minute and it represents the ventilation process. Normal RR values in an adult vary from 12 to 20 breaths per minute [53], [54]. Shifts in this variable are tachypnea (higher values), bradypnea (lower values) or apnea (nonexistent). These are often the first signs of deterioration due to the body's effort to keep adequate oxygen delivery [55]. Also, ventilation depends on the arterial partial pressure of oxygen and on the arterial partial pressure of carbon dioxide. So, hypoxemia and hypercarbia, two unfavorable conditions, can cause a RR increase [54] that can be identified through this variable measurement. Other conditions that can affect RR are, e.g., strokes, fevers, asthma, heart conditions and infections [53]. RR is measured through the observation of the number of times the chest rises/falls, impedance pneumographs, capnographs [56] and can also be estimated through ECG, PPG or accelerometry signals [51].

2.3 ECG monitoring

Heart's contraction, and consequent blood flow to all body tissues, depends on a coordinated transmission of electrical impulses throughout its intrinsic conduction system. This electrical activity generates potential differences at the skin surface that can be measured using electrodes. The typical ECG signal (Figure 2.1), obtained using lead II, is composed of three main waves (P, QRS complex and T) that correspond to the flow of those electrical impulses throughout the heart, as illustrated in Figure 2.1. Abnormalities in these waves shape and duration and in the intervals duration between them are indicative that a complication may be occurring. In fact, many complications were shown to be preceded by changes in this signal hours before their onset [57]–[59]. Proper interpretation of the ECG also allows the detection of arrhythmias [59], which can both be physiological and life-threatening, depending on its type. ECG monitoring, due to this signal's nature and characteristics, must be made continuously and permits the extraction of several different measures, like the QRSa and the RRI.

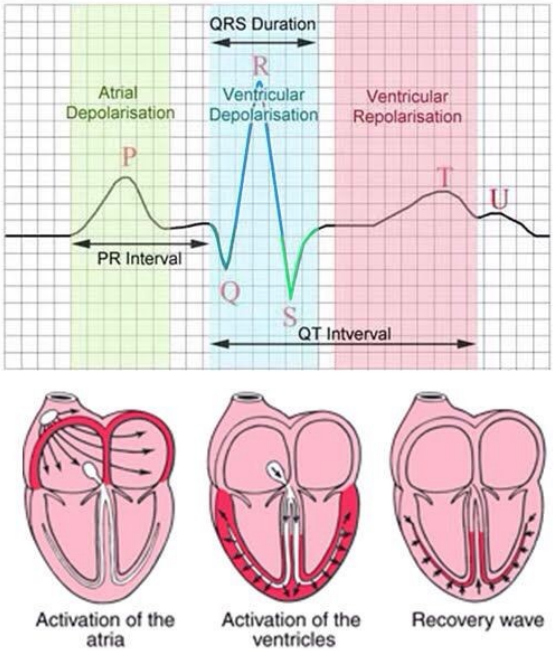


Figure 2.1: Representation of the electrical impulses' propagation through the heart and corresponding ECG signal [60].

2.3.1 QRS complex amplitude

As seen in Figure 2.1, the QRS complex represents the ventricular depolarization, and consequent contraction. Despite mentioning it as QRS complex amplitude (QRSa), in reality, the amplitude of interest is the R-wave amplitude (when using lead II), i.e., the potential, measured in mV with skin electrodes, of the positive wave present in the QRS complex. The three waves that compose a complete QRS complex (Q, R and S) are not always visible and several variations of this complex can occur.

QRSa physiological values range between 0.5 and 3 mV [61]. Yet, a variety of reasons can cause deviations. The distance between the heart and the electrodes is one of them, which means this measure might be very different between slim and obese subjects. Larger amplitudes main originator is the presence of ventricular hypertrophy, since electrical currents generated by ventricular myocardium are proportional to its muscle mass. Other causes include the presence of an abnormal natural pacemaker or ventricular enlargement [62], [63]. Low amplitudes might be prompted by chronic obstructive pulmonary disease, due to thorax hyperinflation, hypothyroidism or pericardial effusion [62], [64].

2.3.2 RR interval (RRI)

RR interval (RRI), also known as interbeat interval, is the time interval between successive R-waves, and it usually stands between 600 and 1200 ms [65]. Fluctuations in its values are a measure of heart rate variability. This variability in a healthy heart is complex and non-linear and provides flexibility to respond to an uncertain and ever-changing

environment [66]. This is controlled by the autonomic nervous system, in particular, by a dynamic balance between parasympathetic and sympathetic activity.

Pathological conditions can either increase or decrease this complexity. Elevated values might occur due to atrial fibrillation. In cases where those values arise due to cardiac conduction irregularities, there's a link with increased risk of mortality, especially in older individuals [66]. On the other hand, a reduction in this variability has been associated with cardiovascular disease, like heart failure [67], and diabetic autonomic neuropathy [68]. More important, this reduction is an indicator of clinical deterioration and its use was suggested for patient monitoring and as assistive tool for decision support [69]. In fact, the analysis of this signal was identified as a method to quantity the risk of developing different arrhythmic events or even death [70].

The bottom line is that this measure enables the assessment of cardiac health, as well as of the state of the autonomic nervous system.

2.4 Machine learning

2.4.1 Fundamentals and notation

Machine Learning (ML) is a subfield of artificial intelligence that seeks to look for patterns in data with the aim of improving a performance criterion to make use of in future decisions, based on previously provided observations/examples [71]. One of the most famous ML definitions is given by Tom Mitchell [72]: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". This is accomplished by programming algorithms that intend to automatically build a computational model of the relations between observable quantities (inputs) and the desired variables (outputs) [73]. Those models main purpose may be to cluster data, reduce dimensionality, make predictions, amongst others. In this project, the interest will primarily pass by prediction models.

Prediction models use previous data to learn how to accurately predict an output variable, and then apply this knowledge to make similar predictions on new data. This is possible by learning a predictor function, that maps input variables to an output one, using statistical techniques [74]. This type of learning is defined as supervised learning and, in this, the output variable (the correct "answer") is known when fitting the model. Supervised learning is generally divided into two main categories:

- **Classification problems** – input variables are mapped to a **discrete** output variable (label).
- **Regression problems** - input variables are mapped to a **continuous** output variable.

The process of developing a prediction model usually consists of four main steps:

- **Data acquisition** – collect a set as large as possible and store it in a suitable way to be computationally processed. Preprocessing of this data might be needed before the next step, depending on the problem.

- **Feature extraction and selection** – selection of which data characteristics represent relevant properties for predicting the output. It involves transforming the data into a different space of variables. To optimally choose which and how many features should be provided to the model, it might be necessary to perform dimensionality reduction. The implementation of feature selection procedures is, perhaps, one of the most typical ways to do that [75]. The result of this stage is a dataset $D := \{(\boldsymbol{X}, \boldsymbol{Y})\}$, where $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ represents the matrix of features and each of its rows, $\boldsymbol{x_i}$, is a row vector of features. m is the total number of data examples/observations and n is the total number of features. $\boldsymbol{Y}$ is a column vector with the same number of rows as $\boldsymbol{X}$, where each element, y_i, represents the output variable (a label for classification problems and a continuous number for regression problems) associated with $\boldsymbol{x_i}$.

- **Learning** – it's hypothesized that there's a true function, f, such that $y_i = f(\boldsymbol{x_i})$, for every possible example. Learning is the process of discovery of an approximate predictor function, $\hat{f}$, using a subset of D, the training set, and a ML algorithm, that performs well on previously observed examples and, more important, on yet unobserved ones. This results in a (mathematical) model that, when fed with the same type of features from a new example, is able to make a prediction about it. Usually, a validation set is used to tune the model's hyperparameters[1]. This set might consist of a portion of the training set held back from training to provide an estimation of the model skill during this hyperparameter tuning. Cross-validation on the training set can be employed instead.

- **Evaluation** – after learning, the resulting model's performance has to be evaluated. This is done by calculating appropriate metrics, which are discussed in subsection 2.4.3 (only classification problems performance metrics are discussed).

2.4.2 Model types

The underlying problem of this thesis is a binary classification problem, since the major goal is to predict whether a patient will deteriorate (positive class, 1) or not (negative class, 0). With that in mind, two ML algorithms, Logistic Regression (LR) (see 2.4.2.1) and Boosted Trees (BT) (see 2.4.2.2), were explored for the development of prediction models. These will function as decision support models, as explained in section 1.2.

Additionally, a **clustering model** was implemented as part of a novel preprocessing technique. The theoretical background associated with this other kind of ML models and the description of the clustering algorithm used is presented in this subsection (see 2.4.2.3).

[1] hyperparameters are algorithm-specific parameters that allow to control the learning process

2.4.2.1 Logistic regression

Logistic Regression (LR) is a particular case of a generalized linear[2] model where the link function is the logistic function, also named sigmoid function. A link function is a function that will transform the value of the linear prediction obtained, that can go from $-\infty$ to $+\infty$, to a well-defined range. In this case, the range is between 0 and 1, which provides a probabilistic interpretation of LR outputs. Particularly, LR models the probability of the positive class as shown in the following equation:

$$P(y_i = 1|\boldsymbol{x_i}) = g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_n x_{in}) \tag{2.1}$$

where $P(y_i = 1|\boldsymbol{x_i})$ represents the probability of the positive class given the input vector of features $\boldsymbol{x_i}$, n is the number of features, x_{ij} represents feature j of the features vector $\boldsymbol{x_i}$, β_j is the regression coefficient associated with feature j, β_0 is the intercept term and $g(\cdot)$ is the link function, which, in the case of logistic regression, is given by:

$$g(t) = \frac{1}{1 + e^{-t}} \tag{2.2}$$

In LR, the predictions are no longer a linear combination of the inputs due to the mentioned transformation. Instead, that linear combination is associated with the log odds[3] of the positive class, given that input, as shown in equation 2.3 (which can be obtained from equation 2.1). Consequently, the regression coefficients can be interpreted as the estimated increase in the log odds of the positive class, per unit increase of the value of the associated feature [76], [77]. This also means that the predictor variables have a linear relationship with the outcome on the log odds scale [78].

$$\ln\left(\frac{P(y_i = 1|\boldsymbol{x_i})}{1 - P(y_i = 1|\boldsymbol{x_i})}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_n x_{in} \tag{2.3}$$

Learning, in a LR model, is finding the best set of parameters $\{\beta_1, \beta_2, ..., \beta_n\}$ through maximum-likelihood estimations, using the training set. The best set of parameters is one that minimizes the errors between the probabilities predicted and the examples true label [76], [78]. The idea behind maximum-likelihood estimations is to maximize a likelihood function towards that minimal error. A likelihood function can be interpreted as a measure of the quality of the model fit to the data. One of these functions, and one that is explored in this thesis is the log-likelihood:

$$ll(\hat{\beta}) = \sum_{i=1}^{m} y_i \ln(P(y_i = 1|\boldsymbol{x_i}, \hat{\beta})) + (1 - y_i)\ln(1 - P(y_i = 1|\boldsymbol{x_i}, \hat{\beta})) \tag{2.4}$$

where $\hat{\beta}$ is the current estimation for the regression coefficients, $ll(\hat{\beta})$ is the log-likelihood with those coefficients, m is the number of examples in the training set, y_i

[2] the inclusion of interaction terms and higher degree variables won't be discussed in this thesis, so it's always implicit that the different features are only linearly combined

[3] odds are the ratio of the probability of an event divided by the probability of the event not occurring, $\frac{p(E)}{1-p(E)}$

is the true label of example i and $P(y_i = 1|\boldsymbol{x_i}, \hat{\beta})$ is the predicted probability of the positive class for example i, given its features vector, $\boldsymbol{x_i}$, and the current estimation for the regression coefficients.

As mentioned before, a LR model predicts a probability. However, it is employed in binary classification problems. This transition from a probability to a discrete binary outcome (0 or 1) is usually performed by defining a threshold that separates both classes. This can be simply made by defining the final prediction $\hat{y_i} = 1, if P(y_i = 1|\boldsymbol{x_i}, \hat{\beta}) \geq 0.5$ or, in more intelligent ways, like choosing the threshold that maximizes a chosen metric, using the validation set.

A noteworthy advantage of LR is that, despite being applied in classification problems, it returns probabilities, which is extremely useful for interpreting how sure the model is about the predictions it's making.

2.4.2.2 Boosted trees

Boosted Trees (BT) is a ML algorithm that combines simple decision trees with a technique called boosting.

Decision trees, also known as Classification and Regression Trees (CART), and all tree-based models, partition the feature space into a set of rectangles and are usually displayed as a sequence of "if-then" statements in the form of a tree [78]. Figure 2.2 highlights the differences between the data partition, in a 2D feature space, that can be achieved when logistic regression or a decision tree are used. In the former, the space is divided by a line, whilst in the latter it's divided by a set of rectangles.

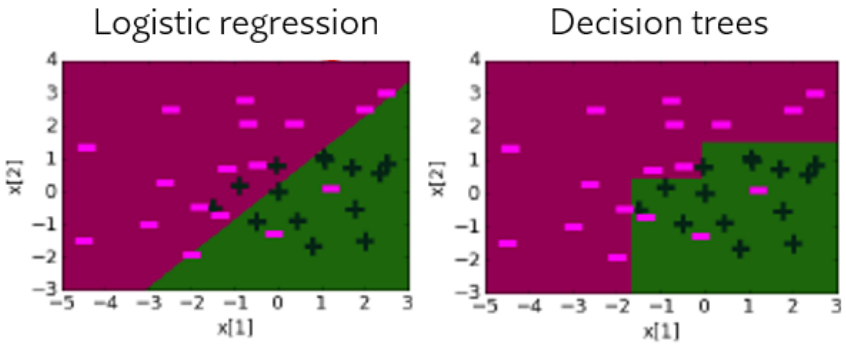


Figure 2.2: Illustration, in a 2D feature space, of possible data partitions using logistic regression (left) and decision trees (right). x[1] and x[2] are features. The pink "minus" symbols represent the negative class examples and the black "plus" symbols represent the positive class examples. The pink region represents the portion of the feature space where examples would be predicted as negatives and the green region represents the portion where examples would be predicted as positives. Adapted from [79].

Decision trees employ recursive greedy algorithms[4] that minimize a cost function

[4]greedy algorithms are algorithms that make the optimal choice at each decision step. This, however,

when selecting which feature and which cut-off point are the optimal data partition at that stage of the tree. Hence, there's no guarantee that the final tree is globally optimal, only that each split was optimal at that point. Also, this algorithm is very sensitive to small changes in the data, making it a very unstable method [78].

Ensemble methods, like bagging (not discussed) or boosting, can be employed to enhance both performance and stability in decision trees. The idea behind boosting started with the work of Kearns and Valiant [80], [81], that posed the "Hypothesis Boosting Problem". In theoretical terms, this hypothesis questions if the notions of weak and strong learnability are equivalent. In practical terms, this can be interpreted as if a set of weak learners[5] can be combined to create a stronger learner. A few years later, Schapire [82] found that this hypothesis was correct and, since then, a wide variety of different boosting techniques have been developed.

Boosting, like all ensemble methods, combines the predictions of multiple, usually weaker, models to produce an ensemble classifier with superior predictive performance. The process of building a binary classification boosted model starts by redefining the classes to be -1 and +1, instead of 0 and 1, as had been reported until now. With this new definitions, the resulting ensemble model predictions can be mathematically expressed as [79]:

$$\hat{y_i} = sign(w_1 f_1(\boldsymbol{x_i}) + w_2 f_2(\boldsymbol{x_i}) + ... + w_T f_T(\boldsymbol{x_i})) \tag{2.5}$$

where $\hat{y_i}$ is the model's prediction for observation i, $f_j(\boldsymbol{x_i})$ is the prediction of the j^{th} weak classifier given the input vector of features $\boldsymbol{x_i}$, w_j is the weight associated with classifier f_j, T is the total number of weak classifiers and $sign(\cdot)$ is the sign function.

With a closer inspection at equation 2.5, it can be concluded that the boosting learning process involves learning each weak classifier, f_j, and its corresponding weight, w_j. In a very simplistic way, this process' steps are:

1. initializing each example in the training set with a weight (do not confuse this weights with the ones mentioned in equation 2.5), usually equal to $1/m_{tr}$, where m_{tr} is the total number of examples in the training set.

2. generating the best classifier (f_j) based on the current weights.

3. calculating w_j accordingly to f_j's performance, i.e., w_j is a higher positive number if f_j performs well and is a lower negative number if f_j performs badly.

4. reweighing the examples by giving more weight to misclassified ones and less weight to correctly classified ones, which will increase the misclassified examples importance.

does not guarantee that the globally optimal solution is produced

[5] a weak learner is an algorithm/classifier whose hypothesis performance is only slightly better than random chance [80]

5. repeating steps 2. to 4. iteratively until a stopping condition is reached (for example, until a predefined number of weak classifiers is trained).

6. combining all learners as in equation 2.5.

This learning process implies that observations that are harder to classify receive increasingly higher weights until a learner manages to correctly classify them. This means that each weak classifier will learn different aspects of the data and will focus on regions with difficult-to-classify observations [83]. Additionally, it results in each learner being dependent on past learners and in an unequal contribution of each of them to the final model.

Depending on the boosting technique, the learning process might be slightly different from the one detailed here, which was mostly based on AdaBoost. This was the first successful boosting technique and it's the one implemented in this thesis. For a more detailed explanation on AdaBoost and other boosting techniques see Kuhn et al. and Friedman et al. [83], [84].

Boosting is a simple and interpretable approach, that has been widely used in the industry [79], with many advantages and given proofs. Decision trees are a perfect type of base learner to combine with boosting since they can easily be restricted to be weak learners, are computationally inexpensive to generate and its predictions are easily added together [83]. These characteristics prompted the use of Boosted Trees (BT) in this thesis.

BT can automatically threshold features across their entire range and explore interactions and non-linear effects between them (as shown in figure 2.2), which must be made manually in other algorithms like LR. Besides that, it can lead to improved performance and grants the opportunity of discovering important interactions and threshold values [78], which gains relevance in a medical context. However, its enhanced flexibility can lead to overfitting situations. In fact, previous studies suggested that these modern and complex techniques require larger datasets to produce a stable model [85].

2.4.2.3 Clustering - K-prototypes

Clustering models intend to find natural groups, or clusters, occurring in the feature space of input data. These apply mostly when there is no label or number to be predicted but rather when the data is to be divided into groups [86]. These models are based on a type of learning designated unsupervised learning and, in this, there's no output variable (no previously known correct "answer") available when fitting the model. Instead, the goal of this type of learning is to find hidden patterns and structure in the data by itself, i.e., without being previously informed of what are the possible and correct "answers" [87]. Nonetheless, if these "answers" are known they can be used to evaluate the model but never to fit it.

The process of developing this type of models is similar to that of prediction models, in terms of which are the main steps that have to be completed. However, there are differences in the way some of these are performed:

- **Feature extraction and selection** - the difference here is that the resulting dataset only contains the matrix of features, since the output variable is unknown.

- **Learning** - here, and in a very summarized way, learning involves the use of similarity or distance measures between examples, to identify distinct and dense regions of observations, the clusters [86].

- **Evaluation** - focusing on cases where there are no ground truth labels, the discussion in subsection 2.4.3 cannot be applied. Instead, clustering quality is usually assessed using measures that analyze intra-cluster and inter-cluster distances, such as the silhouette coefficient or the Calinski-Harabasz index.

Real world datasets often contain both numerical and categorical features. However, most clustering algorithms are limited to deal with either numerical or categorical values, but not both. K-prototypes is a combination of the well-known k-means and k-modes algorithms that supports clustering with these mixed-type variables datasets [88]. It incorporates the clustering process of k-means but it uses the k-modes approach to update the clusters centroids' categorical values. Also, it resorts to a dissimilarity measure that integrates both categorical and numerical attributes:

$$d(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j) \tag{2.6}$$

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \tag{2.7}$$

where X and Y are two different examples/objects described, in this order, by p numerical attributes and $m-p$ categorical attributes and γ allows to adjust the weight that is given to each type of attribute. The first term in equation 2.6 is the squared Euclidean distance applied to the numerical attributes while the second term is a matching dissimilarity measure applied to the categorical attributes and defined in equation 2.7 [88]. A more detailed explanation of the k-prototypes algorithm, as well as of k-means and k-modes can be found in Huang's work [88], [89].

Two remarkable advantages of this algorithm are that it preserves the efficiency of the k-means algorithm and it provides interpretability, i.e., each cluster centroid has meaningful values for both numerical and categorical attributes, providing a conceptual description for each of the identified clusters [88].

Other options to cluster mixed-type data are the use of Gower distance with k-medoids or hierarchical clustering and implementing k-means after one-hot encoding the categorical features. For advantages of k-prototypes when compared with these and other options see Huang [88].

2.4.3 Performance metrics

To properly evaluate a prediction model, it is common practice to split the dataset D into training and test set, which are independent. The first is used to fit the model, as mentioned before, whereas the latter is used to provide an unbiased evaluation of the model [90]. Some splitting methods are n-fold cross validation, split sets, leave-one-out and bootstrapping.

A fundamental concept in classification problems evaluation is the confusion matrix. It provides a tabular view of the model's predictions against the true labels, as illustrated in figure 2.3. Some of the most commonly retrieved metrics from this matrix are accuracy, $\frac{TP+TN}{TP+TN+FP+FN}$, recall or sensitivity, $\frac{TP}{TP+FN}$, precision, $\frac{TP}{TP+FP}$, specificity, $\frac{TN}{TN+FP}$, false positive rate, $\frac{FP}{FP+TN}$ and the F_1-Score, $2 \times \frac{precision \times recall}{precision+recall}$ (all of them vary between 0 and 1). It's worthy to reiterate that all these metrics should be reported considering the test set.

		Predicted class	
		Negative	Positive
True class	Negative	True negatives (TN)	False positives (FP)
	Positive	False negatives (FN)	True positives (TP)

Figure 2.3: Confusion matrix for a binary classification problem.

Another very extensively reported metric is the Area Under the receiver operating characteristic Curve (AUC). This curve is a plot of false positive rate vs recall and shows the performance of a binary classifier across all threshold values (for threshold explanation see 2.4.2.1). The AUC is thus a threshold invariant metric, that combines the model's performance at various thresholds. It ranges from 0 to 1 and higher values indicate a better model at predicting negatives as negatives and positives as positives.

Throughout the thesis, two other performance curves are presented. The EWS efficiency curve is a plot illustrating recall vs the percentage of observations above the respective threshold. In the context of this thesis, this curve enables the visualization of the proportion of alarms that have to be set off to achieve a certain value of recall. The second curve, the precision-recall curve, provides a visualization of the trade-off between precision and recall across all different threshold values. It is very important in problems with an imbalanced dataset and when the concern is higher on the correct prediction of the minority class (deterioration cases in this thesis).

Chapter 3

Literature review

This chapter presents a literature review on recent research regarding the development of decision support models, for the automatic detection of clinical deterioration. First, the current methods used in clinical context are analyzed, followed by an exploration of more recently developed models and novel strategies. Besides academic studies, commercial solutions already on the market are also examined. Aside from the strategy itself, further recommendations found are also introduced in this review.

Section 3.2 summarizes the review's conclusions and presents a compilation of the most relevant recommendations and limitations identified in previous studies, regarding the development of warning systems, which are to be addressed in this thesis.

3.1 Review

The recent appearance of Electronic Medical Records and wearable sensors has allowed the use of demographics, comorbidities, vital signs, laboratory tests and other clinical parameters as tools to automatically detect clinical deterioration. This is accomplished by combining this now digitally available information, with single or multi-parameter scores and machine learning algorithms, to develop warning systems [5].

Early warning scores

Early Warning Scores (EWS) have been the standard method for deterioration detection since they were introduced in medical practice. EWS evaluate vital signs (and other clinical variables) and assign scores accordingly to the severity of deviations from their normal ranges (see table 3.1 for an example) [5]. If the sum of scores surpasses a certain threshold, it's assumed that deterioration is occurring, and an alert is emitted so that

an adequate clinical response can be initiated. The most often cited EWS are MEWS (Modified) [91], which is illustrated in table 3.1, ViEWS (VitalPAC) and NEWS (National). Slight variations might be implemented accordingly to local hospital standard protocols. Single-parameter scores are similar to EWS but the alarms activation only depends on one clinical variable at a time.

Table 3.1: Modified Early Warning Score. The sum of scores from each variable corresponds to the final score, which is then compared with a given threshold. If the final score exceeds that threshold, a deterioration alarm is emitted.

Score	3	2	1	0	1	2	3
Respiratory frequency (breaths/min)		≤8		9-14	15-20	21-29	≥30
Heart rate (beats/min)		≤40	41-50	51-100	101-110	111-129	≥130
Systolic blood pressure (mmHg)	≤70	71-80	81-100	101-200		≥201	
Consciousness			Confused	Alert	Responds to voice	Responds to pain	No response
Body temperature (℃)		≤34.9		35-38.4		≥38.5	
Urine production (ml/hour)		≤20	21-39	≥40			

Another score-based approach is the Rothman Index, which is an index that assesses patient's condition based on 26 variables obtained from its Electronic Medical Record [92]. Some flaws presented by both EWS and the Rothman Index are: (1) not being specific for surgical patients; (2) requiring periodical manual measurements by nurses, which are subjective and affect nurses' workflow and workload (besides other issues stated in section 1.1); (3) requiring data insertion in the hospital management system (only Rothman Index), which is not always promptly entered; (4) lacking personalization, for example, EWS use generic thresholds that are neither adapted to a patient nor a particular condition/population; (5) vital signs are analyzed independently, leaving possible correlations unexplored. Furthermore, Gao et al. [93] performed a review on many simple warning systems, which included score-based methods, like EWS, and reported little evidence of reliability, poor sensitivity and poor predictive value. This highlights the fact that these scores cannot fully cope with the complexity of these patients' physiology [5]. These score-based approaches were also already explored from a different point of view, i.e., it was investigated if the analysis of the postoperative evolution of these scores in time could result in better performance [5]. Despite obtaining promising results, this new procedure suffers from the same previously mentioned score-based approaches' limitations, as it depends on the scores calculation.

Given the inaptitude of score-based approaches, an alternative was required. Therefore, researchers focused their efforts on the exploration of the ML field. In fact, most recent studies either employ the **novelty detection** approach (also known as one-class classification) or the **binary classification** approach (also known as two-class classification), both being ML techniques.

Novelty detection

In this context, novelty detection usually involves developing a model of normality, using data from patients that didn't present complications during their hospital stay, and then adopting distance metrics or statistical methods to evaluate how "abnormal" is a test observation with respect to that model. Therefore, an objective measure of patient's deviation from what would be a healthy state is returned as result of employing a novelty detection model. This approach is often used when data from patients that deteriorated is scarce, as it's much more challenging to model non-normal classes in these conditions [5].

Pimentel et al. [94] took this approach further and also analyzed the vital signs trajectory in time by building a model of normality based on patients' vital signs in the day of discharge, using a kernel density estimate, and then compared it to measurements in other previous periods of ward stay. Significant physiological trajectories differences between patients who developed complications and those who did not were obtained. They also performed a similar study [95] but using a gaussian process regression to model normal vital signs trajectories and were again able to discriminate abnormal trajectories. However, these models didn't explore possible correlations between vital signs, the data wasn't continuously acquired and the models had no degree of personalization.

Additionally, Clifton et al. [96] presented very good results using a one-class support vector machine and a gaussian mixture model on manually acquired vital signs data and on a synthetic dataset. A few years later, the same team [97] compared four novelty detection algorithms, but this time exploring mostly data continuously acquired using wearable sensors. They proved that these methods are suitable to identify deterioration, with high accuracy and sensitivity, and even obtained better specificity than EWS. As future reference they denoted the importance of personalizing the models and suggested that patient-specific data can be sequentially introduced in the previously constructed model, as it is collected.

In fact, in-between those studies, they performed another one [98] that included a personalized method that exploited gaussian processes for inference of periods of missing or erroneous data, which could have arisen due to patient's movement or sensors issues. They showed that deterioration can be predicted earlier by introducing personalization in the models. In this matter, evidence [5] suggests that algorithms should take into consideration the particularities of surgical patients (rather than using fixed general thresholds) and the different baseline and contextual characteristics of each individual.

Indeed, other patient-specific models have already been reported [99], [100] with good predictive capability.

Finally, Tarassenko et al. [101] reviewed a novelty detection system, BioSign, which is based on a combination of k-means clustering and kernel density estimates on continuously acquired data. They reported that 95% of the generated alarms were true. BioSign is now known as Visensia, a commercial available system discussed further ahead in this section.

Binary classification

Despite a number of novelty detection strategies can be found in the literature, the binary classification approach is also a very commonly employed one. It consists in developing a model that, given some set of features, returns a prediction regarding whether an observation belongs to one of two classes. In this context, the approach usually involves providing as input to the model, features extracted from data acquired in a certain past interval, e.g., previous 24 hours, and predict positive class, if a deterioration event is expected to occur in a certain future interval, e.g., the following 8 hours, and negative class, otherwise.

Khalid et al. [102] performed one of the pioneer studies employing this approach, with a ML algorithm called support vector machine. They evaluated the reliability of using class labels, attributed by clinicians, for each set of measurements, and proposed a novel method to automatically refine them. However interesting, this approach relies on the manual labeling by experts, which is a slow, time-consuming and error-prone process.

In a slightly different perspective, Churpek et al. [103] developed a new score based on LR models and on a variety of manually acquired measures, such as vital signs or laboratory results. Their dataset included patients across five different hospitals and all results obtained outperformed the MEWS. Escobar et al. [104] implemented a similar strategy (using LR) but their model had an additional disadvantage of requiring more measures, like care directives, to be collected. They included a wide range of non-ICU patients' populations, which resulted in the model performing very differently for each of them. This highlights that a population specific model, or, at least, a model specific for surgical patients, might lead to enhanced results.

A few years later, in a different study, Churpek et al. [78] revealed that several ML algorithms, such as random forests, can outperform LR and EWS in detecting clinical deterioration. Starting from a similar hypothesis, Pirracchio et al. [105] developed the Super Learner, which is an ensemble nonparametric method for constructing a refined predictive algorithm given a set of candidate ML algorithms. However it obtained very good results, it was tailored for ICU patients and required many variables to be acquired. Dal Canton et al. [106] also explored the potential of a variety of ML algorithms, however with unsatisfactory results.

In a more singular approach, Stevens et al. [107] employed a particular type of ML algorithm, a fuzzy logic classifier. These explore a set of fuzzy rules and fuzzy vital signs thresholds, which contrasts with EWS that are based on exact thresholds. Their system would also differentiate between alarm levels, accordingly to the patient's derangement severity. Additionally, they combined it with a support tool, based on Bayesian theory, that displays possible complications the patient might be undergoing. Using this approach they verified a significant reduction in false alarms but the dataset used was very small and focused on ICU patients.

Mao et al. [108] proposed a solution that not only sought for higher performance but also addressed the class imbalance issue, present in most of the datasets surrounding the problem described in this thesis, with an exploratory undersampling technique.

Finally, Moss et al. [109] ascertained if the inclusion of ECG-based measures could improve the predictive ability of their models. The results were positive, with the inclusion of such predictors showing a consistent improvement in performance.

Noteworthy to mention that a huge majority of the models developed in the reviewed papers, that would compare their results with currently in-practice EWS, obtained far better results.

A summary of the limitations identified in the reviewed work, regarding the development of warning systems based either on binary classification or novelty detection, is provided in the next section (3.2). In addition to that, a review table that summarizes the discussed studies is presented in appendix A.

Commercial solutions

Besides purely academically investigated warning systems, some commercially available ones exist already. Visensia is a software that produces a safety index based on a model of normality derived from a high-risk population on a general ward. It works with both continuous and periodic variables measurements. This is a very expensive software, not specific for surgical patients, that usually requires connection with bedside monitors when used in hospital settings. Complications are predicted with 6 to 10 hours in advance [32]. Visensia has demonstrated better results than current detection techniques in several studies [110], [111].

Sensium is an early warning system for clinical deterioration, designed for ambulatory monitoring of the general ward population. It includes a wearable patch, that provides continuous vital signs monitoring. Also, it updates HR, RR and BTemp measures every 2 minutes and ensures patient's freedom of movement [22]. It was proven to be acceptable and practical to patients [112] and it is the only wearable patch-based vital sign monitor with published data showing clinical and economic benefits [22]. However, this system is not specific for surgical patients and the algorithms incorporated to generate notifications of deterioration are as simple as EWS.

EarlySense is a low-acuity continuous monitoring system that consists of a piezoelectric sensor placed under the patient's mattress, a bedside monitor and a software to analyze and display data. It provides measures of HR, RR and bed motion updated every 0.5 seconds. The deterioration notifications are generated through threshold-based (same complexity as EWS) or trend-based deviations (comparing the median of readings of a certain day period with the same period of the previous day) [113], [114]. Despite having reported promising results in one study [114], it still presents some limitations, like requiring the patient to be in bed, not being specific for surgical patients and presenting an EWS-like deterioration detection algorithm.

This review on commercial systems clarified that no ideal solution is already in place in clinical settings, since most still rely on EWS and/or bedside monitors. Instead, the prefect scenario would be the combination of more advanced algorithms for predicting deterioration events with the continuous monitoring provided by wearable sensors. This would result in a low-cost warning system with enhanced performance.

Recommendations and findings

Besides the approach to take, further recommendations can be found in the literature. Petit et al. [5] advocated that, at least for elective surgery patients, their preoperative baseline vital signs can be integrated in a personalized model. This allows to estimate the patient's normal profile and hence find similar cases from which a recovery pattern can be predicted. They also proposed the inclusion of other pre- and intra-operative measures. Thompson et al. [115] found a dependency between the most typical postoperative complications and the postoperative period in which those occur more frequently. This might be a helpful tool to dynamically adjust the relevance of a certain vital sign deviation, accordingly to what are the most incident complications in the time period at which the patient presents itself. Raymond [36] reviewed several studies and reported that patients presenting comorbidities are in increased risk of developing postoperative complications. Therefore, comorbidities should be included when developing a clinical deterioration detection model. Also, Batchinsky et al. [33], [69] showed that advanced RRI measures can improve the effectiveness of predicting instability. Cuthbertson et al. [116] were able to detect differences between patients that were readmitted to ICU and those who didn't, using only HR and RR, with 6 to 8 hours in advance and derived a discriminant function that could detect those differences 48 hours before the readmission. Churpek et al. [78] found that the most important variables for their algorithm were major predictors in prior research (RR, HR, age, ...). This inspection of which features contribute more to the model decision might lessen the clinicians' suspicion surrounding ML algorithms and offers explainability to the model. Fieselmann et al. [117] also reported that RR was a better predictor of cardiopulmonary arrest than HR or BP. Cretikos et al. [54] found numerous papers that stressed that RR changes can anticipate deterioration earlier and can identify at-risk patients 24 hours before the event. Chen et al. [118] performed a

study to discriminate between real and false alerts, and reported a vast list of vital signs-based features that can be implemented in order to attain a better differentiation. Kellett et al. [119] developed the Simple Clinical Score, which was based on predictors identified by a LR model and obtained at time of admission only. It intends to stratify patients accordingly to their risk of death within 30 days of admission to acute care settings. Risk scores like this might be an important complementary information for nurses to prioritize treatment and could also be included in the developed models.

These were some of the recommendations found that should be considered when developing clinical deterioration detection models, since their examination can help achieve superior performance and better patient outcomes. A compilation of recommendations is provided in the next section (3.2).

3.2 Conclusions

EWS are the standard practice for the task of assisting nurses detecting clinical deterioration events in hospitalized patients. The inaptitude and constraints presented by this method led researchers to look for new strategies. The solutions encountered combine intermittent and/or continuous monitoring with one of three strategies to interpret and analyze the acquired data: EWS-like models, novelty detection and ML-based binary classification. However much work in the area can be found in the literature, this review highlighted that no optimal solution was already developed, both in an academical and commercial perspective.

Most commercial solutions available are still dependent on EWS-like methods. Additionally, and despite these solutions already involving continuous monitoring, it is yet mainly performed resorting to bedside monitors. These two factors limit both patient monitoring quality and quantity.

Regarding the academical work reviewed, a summary table was produced (see appendix A), which is unique in the topics chosen to be summarized. These enable the reader to compare the reviewed studies on a model building perspective (prediction strategy, train/test partition, model approach, ...) and to focus on the major limiting factors (type of monitoring, vital signs analysis, personalization, ...). One of the main conclusions to withdraw from this table is that no solution was simultaneously based on advanced algorithms to predict deterioration, and on continuous monitoring using wearable sensors alone (independent of manual measurements or bedside monitors).

Aside from the main limitations detailed in the table, more were identified. A complete list of the limitations, presented both by EWS and/or reviewed ML-based models, is now presented:

- the vast majority of solutions are still based on intermittent and manual monitoring. Also, from those who employ continuous monitoring, a relevant portion still uses bedside monitors instead of wearable sensors.

- most of the reviewed studies lacked personalization in the models.
- not being specific for surgical patients in wards. This is rather important due to the very particular characteristics of these patients and to the monitoring conditions in such wards.
- lack of exploration of correlations between vital signs, since they are usually analyzed independently.
- very few outcomes included. Most would only act upon situations that require transfer to ICU or that would result in death.
- poor performance, in terms of predictive value, recall and false alarm rate, which might lead to alarm fatigue in clinical staff.

Besides addressing the above-mentioned limitations, the analysis of recommendations and findings revealed in previous studies can also contribute to the development of a system with higher predictive capability. A summary of recommendations is now presented:

- include pre-, intra- and postoperative predictors.
- include features extracted from major predictors found in prior research, like RR.
- include heart rate variability measures.
- the list of predictors should also include comorbidities and other demographic and contextual features.
- the model should be explainable (not a "black box") and guide clinicians on which features are motivating the alarm.
- the system should predict deterioration early enough to ensure patients can be treated in time.
- the system should fundamentally be dependent on automated measures and should be as simple as possible. This simplicity requirement includes not only the prediction model itself, but the amount of measures that have to be collected as well.
- dynamically change features relevance accordingly to the time-period of patient's stay, driven by which complications usually prevail in that period.
- improve current preprocessing techniques, solutions to deal with missing data and techniques to address class imbalance [5].
- always have in mind the model's computational costs because it still has to be applied with an adequate frequency in real clinical contexts.

Given these lists, the intent of the work performed in this dissertation can be described as the development of a high-performance warning system that tackles as much limitations as possible and considers as much recommendations as possible. Because of the advantages mentioned in the present chapter and in chapter 1, this system should be based on continuous monitoring provided by wearable sensors and on advanced algorithms to generate deterioration alarms. This is the combination that seems to have more advantages in practical terms and that looks more promising to achieve good performance in the deterioration prediction task.

Chapter 4

Materials and dataset

The data explored in this dissertation had been previously acquired in a study called "MoViSign: Mobile Vital Sign tracking in high risk surgical ward patients".

Therefore, this chapter starts by detailing the characteristics of the wearable sensors used in that study. This is followed by a short overview on the MoViSign study, the dataset description and a characterization of the study's population.

4.1 Materials

The work discussed in the next two chapters (5 and 6) was mostly implemented using MATLAB (version 2020a for Windows) and the necessary associated packages. Python programming language was also utilized.

The continuously acquired data explored in this study was obtained using three wearable sensors, which are described in the following subsection.

4.1.1 Sensors

Unlike bedside monitors, the use of wearable sensors confers patients with freedom of movement and it is a lower-cost solution for continuously monitoring hospitalized patients. This technology can have greater impact in general wards, where there is a higher number of patients. With that in mind, three wearable sensors were used to acquire the physiological data explored in this thesis.

Isansys LifeTouch

This sensor was manufactured by Isansys and it's a small light-weight wireless wearable device that continuously acquires the ECG trace, through a pair of electrodes attached to a patient's chest (Figure 4.1 (a)), with a sampling rate of 1000 Hz and a 12-bit resolution.

Besides the ECG itself, it provides measures of RR, HR, QRSa and RRI. Its battery can last between 4 and 5 days. Additionally, accelerometry data is also acquired, which allows to estimate patient's activity levels [27], [120].

Isansys LifeTemp

This sensor was also manufactured by Isansys and it continuously measures BTemp. It is designed for placement in the armpit and its physical characteristics are similar to LifeTouch (Figure 4.1 (b)). It takes readings every 10 seconds and updates information every minute. Its battery can last for 10 days [27], [120].

Nonin 3150 WristOx2

This is a small portable wrist-worn sensor manufactured by Nonin with a finger probe that continuously measures SpO_2 and the PPG trace (Figure 4.1 (c)). By default, it features a 4 second sampling rate and 1080 hours of memory, although even better sampling rates are available. Its battery can last for 2 days [120], [121].

(a) Isansys LifeTouch. (b) Isansys LifeTemp. (c) Nonin 3150 WristOx2.

Figure 4.1: Wearable sensors used and respective adequate placement.

Sensors integration

Data from the above-mentioned sensors is encrypted and continuously transmitted via Bluetooth to a gateway in the Isansys Patient Status Engine (dedicated tablet), where it can be displayed to the caregivers. Then, it is automatically transferred to a local server within the hospital via wi-fi or 3G/4G, where it'll be stored in a secure database (Figure 4.2). At this point, the data is in conditions to be provided as input for the deterioration detection system, where it'll go through the process described in subsection 6.5.1. The corresponding feedback from the warning system is then sent back to the tablet and/or nurses. This automatic transfer of data avoids transcription errors and eases off nurses' workflow [120], [122].

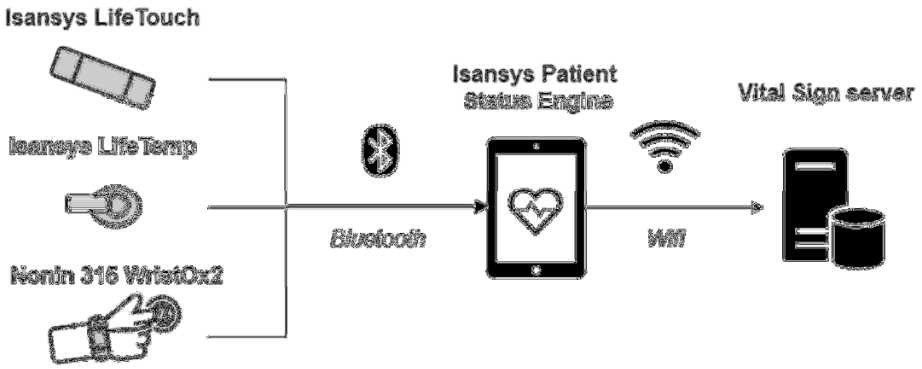


Figure 4.2: Overview of the sensors integration and data transmission.

4.2 Data acquisition

4.2.1 MoViSign study

The dataset explored during this project had already been collected by the research group[1] in a research protocol called "MoViSign: Mobile Vital Sign tracking in high risk surgical ward patients". In this, 60 surgical patients from the ZGT (Ziekenhuisgroep Twente) hospital were continuously monitored with the sensors described in 4.1.1, while recovering in the general ward. Besides that, they received standard ward care, including routine periodical vital signs and MEWS measurements performed by nurses. The study's main goal was to explore to what extent continuous mobile vital signs monitoring could improve clinical deterioration detection, in surgical ward patients, as compared with current methods using MEWS and nurses observations.

During the acquisitions, two EWS, MEWS (Table 3.1) and a novel CEWS (Continuous, Table 4.1), and complications detection by nurse's observations were tested. Event detection by the MEWS was defined as MEWS ≥ 3, by the CEWS was defined as CEWS ≥ 3 for at least 5 minutes in a 10-minutes time frame and by nurse's observations was defined as the first notification of nurse-worry. The results obtained with CEWS were not satisfactory, which enhanced the importance of developing a more adequate decision support model. In order to achieve that, the MoViSign study's resulting dataset was stored and will now be described in more detail.

4.2.2 Data usage and dataset description

The explored dataset was composed by three types of measures/information for each subject:

- continuously acquired - ECG, PPG, **HR** (bpm), **RR** (breaths/min), **BTemp** (°C), **SpO_2** (%), **QRSa** (mV), **RRI** (ms) and activity levels.

[1] Biomedical Signals and Systems group, University of Twente

Table 4.1: Continuous Early Warning Score. The sum of scores from each variable corresponds to the final score, which is then compared with a given threshold. If the final score exceeds that threshold for at least 5 minutes in a 10-minutes time frame, a deterioration alarm is emitted.

Score	3	2	1	0	1	2	3
Respiratory frequency (breaths/min)		≤8		9-14	15-20	21-29	≥30
Heart rate (beats/min)		≤40	41-50	51-100	101-110	111-129	≥130
Skin temperature (°C)		≤34.9		35-38.4		≥38.5	
Oxygen saturation (%SpO2)	≤91	92-93	94-95	≥96			

- periodically acquired - BP, routine manual vital signs and MEWS measurements performed by nurses, laboratory tests results, unplanned interventions information, diagnostic tests effectuated, drug use, pulse regularity, oxygen administration levels, consciousness score and urine production.

- registered once (at time of admission) - **age**, **gender**, **height**, **weight**, **American Society of Anesthesiologists class (ASA)**, **type of surgery**, **number of comorbidities** and **types of comorbidities** (cardiac, vascular, diabetes, pulmonal, neurological/psychiatric, gastrointestinal, urogenital, thrombotic, neuromuscular, endocrine, infection diseases, others).

However, no periodically acquired information was considered, since there was the desire to make the system fully independent of manual periodical measurements. This way, it can repeatedly assess patient's state without requiring any input from nurses or other staff. In fact, the only manually obtained information the system requires are the demographic and contextual measures obtained at time of admission. Regarding the continuously acquired variables, the activity levels, ECG and PPG data were also not used. In the first case, this was a result of the poor data quality, which was consistent across many subjects. In the latter two cases, it was deemed that enough measures extracted from these signals were already available.

So, in summary, only the boldfaced variables were actually utilized. HR, RR, BTemp and SpO_2 measures are available at 1-minute intervals, while QRSa and RRI measures are available every time a heartbeat occurred. This means these last two variables are unevenly sampled.

The dataset is composed of 8916.4 recording hours for each vital sign. However, for some subjects, these included periods in the beginning and in the end of the monitoring period where no actual valid measure was being made. Hence, the total sum of corrected

valid monitoring periods consists of 7883.2 recording hours for each vital sign. The corresponding values for both QRSa and RRI are 7955.1 hours and 7504.1 hours.

During the aggregate of all patient's monitoring periods, 19 deterioration events of interest (see section 2.1) occurred across 16 patients. The list of these events can be found in table 4.2. All the events took place in the postoperative period of the respective patient's stay. This was, however, expected since most complications regarding surgical patients arise postoperatively, as mentioned in section 1.1. From the 19 events, 8 were excluded. 4 because the event timing was missing, 1 because the subject was missing QRSa and RRI data, 1 because the patient had less than 24 hours of data available, 1 because it was the second event springing from the same patient and 1 because the patient's data did not pass an acceptance criterion explained in section 6.1.

Table 4.2: Clinical deterioration events detected and respective number of occurrences, during the study's monitoring period.

Event	**Number of occurrences**
Pneumonia	5
Atrial fibrillation	3
ICU readmission	3
Anastomotic leak	1
Closed loop bowel obstruction	1
Congestive heart failure	1
Death	1
Myocardial ischemia	1
Pleural empyema	1
Urinary tract infection	1
Wound leakage	1
Total	19

So, from the initial set of 60 patients, two separate groups can now be distinguished. The first is composed by the 16 patients that presented at least one deterioration event ("Event" group), whilst the second includes the remaining 44 patients that had a normal and healthy recovery without deterioration events ("Non-Event" group). Nonetheless, and depending on the situation, some of them were excluded from certain parts of the study. Table 4.3 provides a clearer explanation on this exclusion process.

In summary, only 50 subjects (11 "Event" and 39 "Non-Event") were effectively included in the processes of extracting features and developing a prediction model.

4.2.3 Study population

The population included in the MoViSign study was composed by patients aged >18 years, undergoing elective esophageal or gastric resection, admitted to the surgical ward for postoperative care, and patients aged >70 years, undergoing hip fracture surgery, and admitted to the surgical ward for pre- or postoperative care.

Table 4.3: Summary of subjects utilization and respective reasons for exclusion. Situation 1 refers to the subjects used for the development of a clustering model discussed in 5.1.2.1. Situation 2 refers to the subjects that were effectively used for the main branch of this work, i.e, in all steps of the development of the prediction model. Situation 3 refers to the subjects used for implementing the reference EWS discussed in 6.4.1.3.

Situation	**Number of "Event" subjects not included/included**	**Number of "Non-Event" subjects not included/included**	**Reason for exclusion (number of patients excluded for this reason)**
1. New approach - clustering model	1/15	7/37	• Height and weight were missing (8)
2. Prediction model development	5/11	5/39	• Event timing was missing (2) • QRSa and RRI data was missing (1) • Less than 24 hours of data available (1) • Patient's data did not pass an acceptance criterion explained in subsection 6.1 (6)
3. MEWS implementation	3/13	4/40	• Event timing was missing (2) • Patient's data did not have a single observation with at least 4 variables necessary for calculating MEWS (5)

From the initial set of 60 participating patients, only 50 ended up contributing for the accomplishment of the core goal of this dissertation, as mentioned before. Hence, table 4.4 only describes this portion of the initial population. Despite no significant differences were found between the two groups ("Event" and "Non-Event") in any of the analyzed variables, some remarks can be made. First, the "Non-Event" group had a lower mean for the monitoring periods duration, which was expected since these subjects probably also had a shorter hospital stay. Second, the group of patients that underwent esophageal or gastric resection had a higher deterioration rate than those who underwent hip fracture surgery. Finally, it was expected that the "Event" group would present a higher number of comorbidities and an older population [36], which did not happen. This might be related with the small dataset size, since the differences were not significant.

Table 4.4: Comparison between patients in the "Event" and "Non-Event" groups. This comparison includes demographic variables, contextual factors and the monitoring periods duration. The existence of significant differences between the two groups was assessed.

	"Event" group	"Non-Event" group	p-value[a]
Number of patients	11	39	—
Age, years (mean ± SD)	70 ± 12	73 ± 13	0.60[*]
Gender, male	7 (63.6%)	14 (35.9%)	0.11[*]
Type of surgery, gastroesophageal cancer resection	8 (72.7%)	18 (46.2%)	0.13[*]
Number of comorbidities, median (first quartile / third quartile)	3.0 (2.0/4.0)	3.0 (2.0/4.5)	0.91[*]
Predominant comorbidity (% of subjects)	Vascular and gastrointestinal (54.5%)	Vascular (48.7%)	—
Corrected[b] vital signs monitoring period, hours (mean ± SD)	160 ± 80	123 ± 50	0.18[*]
Corrected[b] QRSa and RRI monitoring period, hours (mean ± SD)	160 ± 80	116 ± 51	0.10[*]

SD - standard deviation, QRSa - QRS complex amplitude, RRI - RR interval.
[a] significance assessed using the Wilcoxon rank sum test at 5% significance level.
[b] excluding periods in the beginning and in the end of the monitoring period where no actual valid measure was being made.
[*] not significant.

Chapter 5

Preprocessing

After acquiring data, it generally needs to be preprocessed before anything else. This is a particularly common and important step in ML applications and in situations where data is continuously collected with wearable sensors. The reason for this is that a variety of situations, like sensor detachment, treatment interventions, patient movement or communication issues, can produce segments of erroneous or missing data. Furthermore, preprocessing techniques become crucial in a medical context, since acquisition procedures are designed to enhance the patient's care experience, not to ease a posterior data analysis [123]. These facts prompted the implementation of suitable preprocessing techniques in this thesis. The methodology employed for those techniques implementation is presented in this chapter. This includes the methodology applied for the development of (1) a new approach for handling missing data in vital signs; (2) a novel RRI preprocessing technique. Additionally, related results are presented and discussed.

The four continuously acquired vital signs (HR, RR, BTemp and SpO_2) are generally preprocessed using the same methods, so the same procedure was implemented for all of them. However, the QRSa and RRI signals each requires different strategies. Hence, those are discussed separately.

5.1 Vital signs

The strategy to adequately preprocess the vital signs time series comprehends two stages: **artifact removal**, to deal with periods of erroneous or unreliable data, and **missing data handling**, to deal with periods of absence of data.

5.1.1 Artifact removal

This stage's first step was the automatic removal of obvious outliers. This was achieved by applying physiological thresholding to each of the vital signs time series, according to table 5.1. If any sample was below the lower threshold or above the upper threshold, it would be excluded and replaced as a missing value.

Table 5.1: Physiological thresholds applied to the vital signs time series, so that obvious outliers would be removed. If any sample was below the lower threshold or above the upper threshold, it was excluded and replaced as a missing value.

Vital sign	HR, bpm	RR, breaths/min	BTemp, °C	SpO_2, %
Lower threshold	30	5	30	70
Upper threshold	200	50	50	100

Then, median filtering was applied. This is a technique considered to be suitable to eliminate the short-term variability present in wearable sensors data, since it has been previously used to reduce high frequency noise in physiological data [25], [124], [125]. A 4 minutes window-based median filter was used, as in prior research [125].

Additionally, the BTemp time series required an extra preprocessing step, as advised by Stuiver et al. [126]. They found the BTemp sensor measures accuracy to be outside clinical acceptable limits, thus concluding the sensor to be unreliable. However, so that this sensor's information could still be used, the samples deemed to be unreliable were replaced as missing values. These were the ones that (1) represented a decrease of 0.3 °C in BTemp, comparing with the previous sample, and (2) represented a decrease in BTemp for at least three consecutive samples.

5.1.2 Handling missing data

Several distinct methods to handle missing data periods in vital signs time series have been used before, as illustrated in table A.1. For the most part, these are simple statistical techniques that resort to data nearby the missing period, in order to replace that gap with meaningful values. Some of these approaches are: imputing the last value before the gap (or zero-order hold) [102], [108], linear interpolation [124] and imputing the median or average over previous values [97], [101], [108], [127], e.g., median/average over the last hour of data before the gap.

Besides these simple methods, more complex ones have already been developed. For instance, Sow et al. [125] implemented a recursive application of first-order fading-memory polynomial filters, which revealed unstable for larger gaps, when applied in time-domain. However, a frequency-domain alternative showed a considerable gain in forecasting accuracy. Other approaches involved using gaussian processes [98], [128], where the main associated limitation is their high computational costs [128]. Although

already existing, these more accurate and adequate strategies are yet insufficiently investigated. In addition to that, simple imputations, like averages, still remain the most commonly employed ones [123]. These two facts prompted the development of a new personalized approach during this study, which has two versions.

5.1.2.1 New approach

Background

The inspiration for this new approach came from a previously implemented solution [127]. In this, Sun et al. [127] started by developing a similarity metric between subjects, based on correlations between vital signs and labels provided by experts. At query time, i.e., when a subject's (query patient) data required to have a period of missing data handled, they used its most recent window of available data (assessment window) and the developed similarity metric, to retrieve a set of similar patients. Then, for each of those patients, they would employ a sliding-window approach to identify the data window that best matched the query patient's assessment window, once again, based on the developed similarity metric. Finally, those windows were used to build a regression model that would capture the relations between samples from the query patient and samples from each of the similar patients. This regression model could then be adopted, in conjunction with the data immediately succeeding each similar patient's window, to estimate the missing samples values in the query patient's data.

The main differences between their approach and the one developed in this thesis are (1) they would collect similar patients through a supervised metric learning (based on correlations between vital signs and labels provided by experts), while here it was done with an unsupervised learning-based clustering model (using demographic and contextual features); (2) they used the developed metric to identify the windows that best matched the assessment window, while here it was done with Mahalanobis distance[1]; (3) here, the process of identifying the best matching windows has some additional steps, which are explained further ahead; (4) here, two additional preprocessing steps were included in the last stage of the strategy.

Given the above explanations on Sun's [127] strategy and on the differences between that one and the one developed in this work, the latter's implementation details can now be described.

Implementation

The new approach's goal is to accurately estimate values for missing data periods in a patient's vital sign time series. That patient will from now on be designated the query

[1] $d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T S^{-1}(\boldsymbol{x}-\boldsymbol{y})}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are feature vectors representing the windows, S is the covariance matrix and $d(\boldsymbol{x},\boldsymbol{y})$ is the resulting Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{y}$

patient. This technique can be applied to preprocess its four continuously acquired vital signs (HR, RR, BTemp and SpO_2). The technique's algorithm can be broke down into two major parts: **identifying a set of patients similar to the query patient** and **estimating the missing samples values**, which is dependent on those similar patients' data.

In this context, a patient is said to be similar to other based on their demographic and contextual information. This is, a clustering model was developed to group patients according to their demographic and contextual features: age, body mass index, ASA, number of comorbidities, group (type of surgery), gender and presence or not of certain types of comorbidities (11 types). The number of clusters and the cluster each patient belongs to were unknown *a priori*. In fact, the clustering model's goal is exactly to optimally obtain and report that partition. In its development, the first four features were treated as numerical, while the remaining ones were treated as categorical. All numerical features were rescaled, through z-score standardization[2], before imputed to the model. This is a particularly fundamental procedure when developing clustering models, since these use distance metrics, and each feature, before rescaling, has its own unit of measure, which might have very distinct ranges from feature to feature. In summary, the dataset used to yield the clustering model was composed by 52 instances/patients (see table 4.3), each of them having 17 associated features (4 numerical and 13 categorical).

Since the set of features includes numerical and categorical features, the employed solution for clustering should deal with both types adequately. That being said, the chosen clustering algorithm was k-prototypes, given its advantages facing other methods that also support mixed-type features [88]. The k-prototypes algorithm was implemented using the *kmodes* Python package, and during the implementation, several issues had to be addressed. In particular:

- **centroid's initialization** - the k-prototypes algorithm requires the initial cluster centroids to be specified. For attributing the categorical features values to the initial centroids, a procedure described by Huang [88] was employed, while for the numerical features, the attributed value would be a random value extracted from a normal distribution (where its parameters were the feature mean and standard deviation in the dataset). This selection method produces initial centroids more diverse, which can enhance cluster quality [88].

- **choosing the optimal number of clusters and assessing clustering quality** - evaluating the performance of clustering models is not as simple as evaluating prediction models. This is the case because, in the former, performance is measured by assessing how good was the data separation provided by the clustering, instead of just involving the comparison of predictions with ground truth labels. However, currently, there's already a variety of indices that can be calculated to validate the

[2] $z = \frac{x-\mu}{\sigma}$, where x is the feature value before rescaling, z is the feature value after rescaling, μ is the feature mean value in the dataset and σ is the feature standard deviation in the dataset. The number of subjects included for the calculation of μ and σ are specified in table 4.3, situation 1.

results obtained from clustering. One of which, and the one implemented in this study, is the silhouette coefficient, which has been reported as one of the most reliable ones [129]. This is a measure of how well-matched each instance is to its own cluster, when comparing with the next closest cluster, as described by:

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \tag{5.1}$$

where s_i is the silhouette coefficient for instance i, a_i is the mean distance between instance i and all other data points belonging to the same cluster and b_i is the mean distance between instance i and all other data points belonging to the next closest cluster. All distances are calculated as in equation 2.6.

By averaging this coefficient across all instances in the dataset, a measure of overall clustering quality is obtained [130]. With a closer inspection on equation 5.1, it can be concluded that a s_i value close to 1 indicates that the instance was assigned to the correct cluster, whilst a value close to -1 indicates the opposite. Consequently, an average silhouette coefficient close to 1 implies an efficient clustering, whilst a value close to -1 implies the opposite. This coefficient can then be used iteratively to evaluate the clustering quality of models with different number of clusters. In this thesis, the number of clusters was varied between 2 and 11. The configuration that yielded a higher average silhouette coefficient was considered the optimal one and the corresponding value for the number of clusters was the one used for the final model.

- **gamma parameter selection** - as can be perceived by inspecting equation 2.6, the value of parameter γ must be specified. According to Huang [89], a suitable value for this parameter lies between $\frac{1}{3}\sigma_{avg}$ and $\frac{2}{3}\sigma_{avg}$, where σ_{avg} is the average standard deviation of numeric features. Therefore, the value of $\frac{1}{2}\sigma_{avg}$ was chosen.

- **feature selection procedures** - these procedures can both contribute to reduce the model's computational costs and to improve performance, since features that don't contain useful information can hinder the clustering process. The procedure implemented here was based on mutual information calculation. Mutual information can be described as the application of information gain to the task of feature selection. It is a non-negative entropy-based measure that calculates the dependency between variables, by assessing the reduction in uncertainty for one variable, caused by knowing values from the other. This assessment relies on entropy estimations using k-nearest neighbors distances. A value of 3 for k was used, as recommended in prior research [131], [132]. More details about this measure's calculation can be found elsewhere [131], [132]. A mutual information value of 0 indicates independence between two variables, while the higher the value, the higher the dependency between them. Thereby, mutual information can be employed to evaluate each feature's contribution to the clustering results, by determining the final clusters constitution

dependency on each feature. 95% confidence intervals were calculated, since the mutual information implementation used introduces random noise to numerical variables.

Additionally, in order to check for the presence of redundant features, correlations between them were assessed. Since there were two types of features and three possible combinations between them (numerical-numerical, numerical-categorical and categorical-categorical), different correlation coefficients had to be implemented. In fact, two different correlation coefficients were implemented for each combination. The numerical-numerical correlations were evaluated through the Pearson correlation coefficient and the Spearman rank correlation coefficient (both range from -1 to 1). The numerical-categorical correlations were evaluated through the Kruskal Wallis H test and the eta correlation coefficient (range from 0 to 1). The categorical-categorical correlations were evaluated through the Cramer's V coefficient and the Theil's U coefficient (both range from 0 to 1). Statistical significance was assessed at 5% significance level and using either chi-square or t-test statistics, depending on the coefficient. When using the coefficients with a defined range (all but the Kruskal Wallis H test), a pair of features was considered to be correlated if the corresponding coefficient absolute value was higher than 0.7 and statistically significant. For the Kruskal Wallis H test, it was considered enough to be statistically significant, since no previously reported threshold was found.

That being said, the methodology behind the correlation assessment process was implementing the above-mentioned correlation coefficients, and then, if a pair of features was deemed to be correlated by both the coefficients suitable for the features types, the feature that contributed less to the clustering (measured by mutual information) was pondered to be removed.

Considering both the results from the correlations assessment and mutual information calculations, several features combinations were tested. The combination that yielded a higher silhouette coefficient was considered the optimal one and the corresponding set of features was the one used for the final model.

- **global optimization problem** - since the algorithm doesn't guarantee the globally optimal solution to be achieved [89], several iterations, with different initial centroids, were run. The configuration that would present better clustering quality was the one that was kept. However usually improving the results, this still doesn't assure that the globally optimal solution was found.

After addressing the above-described topics, a final clustering model was obtained. Therefore, at this point, it was already possible to fulfill the first part of this new approach: identifying a set of patients similar to the query patient. This is performed by first extracting the query patient's demographic and contextual features required, followed by the standardization of the numerical ones. Then, these are provided as input for the

clustering model, which will return the cluster the query patient belongs to and its 10 closest patients belonging to that same cluster. This process is illustrated in figure 5.1. As can be ascertained, this process resorts to the previously developed clustering model. In the final step of the process, the distance metric used to identify the closest patients is, once again, the one defined in equation 2.6. 10 patients were chosen to be retrieved based on the number of patients each final cluster had and on an experiment performed by Sun et al. [127] using their strategy. Aside from the same number of identified similar patients, this first part of the new approach has nothing in common with Sun's [127] strategy.

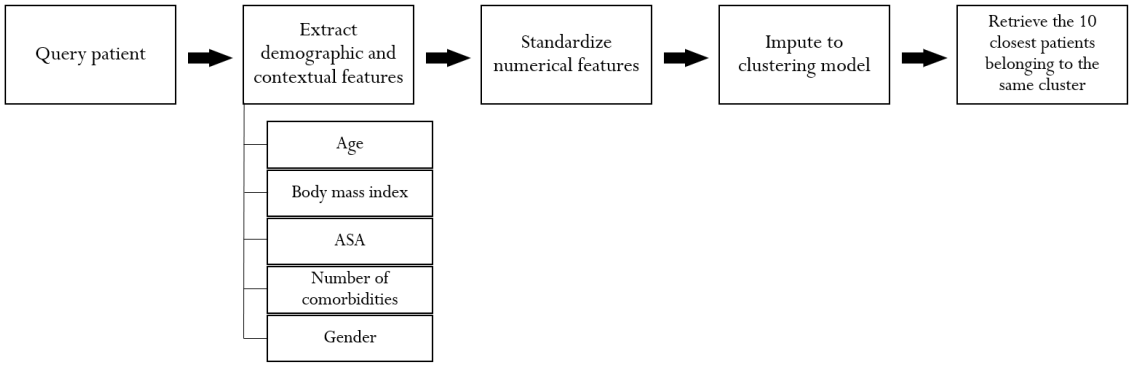


Figure 5.1: Flowchart illustrating the implemented process to obtain a set of patients similar to the query patient.

Before moving on for the second part of the new approach, it's important to mention that the first part only has to be done once for each patient, even though the four vital signs are preprocessed. However, the second part, which is explained promptly, must be executed for every missing data period.

That being said, and having a set of similar patients identified, the second part of the new approach could now begin: estimating the missing samples values. This second part's process comprises the following steps:

- **extracting features from the query patient's previous 60 minutes of data** - this approach requires that the 60 minutes of data before the missing data period do not have any missing values. This segment of data will, from now on, be referred as the assessment window, and it'll be considered that the missing period being handled has a duration of N samples. From the assessment window, 12 features are extracted: mean, standard deviation and the top-10 coefficients of the discrete wavelet transform using the Daubechies-4 wavelet [127]. This set of measures will be referred as the featured-assessment window.

- **finding a window similar to the assessment window, in each of the similar patients data** - for each of the 10 similar patients identified in the first part of this approach, the goal is to retrieve the 60 minutes window of their data most similar

to the assessment window. To achieve that, a course of actions is now explained, which is performed for each of the 10 similar patients.

First, a 60-minutes long sliding-window approach, with 30 minutes overlap, is applied to the patients' data already preprocessed with the median filter (the one used to preprocess all vital signs). This sliding-window approach's purpose is to identify every window that do not contain missing values and, then, extract from them the same 12 features as for the assessment window. This set of feature-extracted windows is compared with the featured-assessment window, using Mahalanobis distance, and the 5 closest ones are kept. Then, orderly from the closest one to the furthest, the N samples immediately after the window were checked. If at least $\frac{N}{2}$ were not missing, both the window and the N samples immediately after the window were kept, and the remaining windows were discarded. If this did not happen for any of the 5 windows, this patient would be removed for the next steps of the approach. The possible missing samples, present in the N samples immediately after the window, would be replaced by the patient mean in the remaining available ones.

To summarize, at this point, a maximum of 10 60-minutes long windows (plus the N samples immediately after) were kept, one for each similar patient. These correspond to their segment of data most similar to the assessment window.

- **developing a linear regression model** - from the windows obtained in the previous step and the assessment window, the dataset $D := \{(\boldsymbol{X}, \boldsymbol{Y})\}$ was constructed. $\boldsymbol{X} \in \mathbb{R}^{60\times n}$ is the set of similar windows, where n is the number of similar patients being considered and must satisfy $n \leq 10$. $\boldsymbol{X}$ has 60 rows since these are 60-minutes long windows and the vital signs were sampled at 1-minute intervals. $\boldsymbol{Y} \in \mathbb{R}^{60\times 1}$ is the assessment window. This dataset was then used to learn a linear regression model, which would capture the relations between samples from the query patient and samples from each of the similar patients:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{5.2}$$

where y represents the query patient's sample, x_i represents the i^{th} similar patient's sample, β_i is the regression coefficient associated with the i^{th} similar patient's samples and n is the number of similar patients being considered and must satisfy $n \leq 10$.

- **estimating the missing samples values** - with the developed linear regression model and the N samples immediately after each similar patient's window, which were previously kept, the query patient's missing values can now be estimated:

$$\hat{y_j} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_n x_{nj} \tag{5.3}$$

where $\hat{y}_j$ is the estimation for missing sample j, x_{ij} is the j^{th} sample immediately after similar patient i's window, β_i is the i^{th} regression coefficient learned in equation 5.2, n has the same meaning as in equation 5.2 and $j = 1, 2, ..., N$.

- **median filtering** - after the N samples were estimated, the resulting new segment of data is preprocessed with the same median filter used before to preprocess the vital signs time series. This is the last step of the new approach version 1.

- **thresholding (only version 2)** - this additional step consists of, first, checking the last available sample before the gap and the first available sample after the gap. The one with higher value is assigned as *max* and the other as *min*. Finally, all newly estimated and preprocessed samples, replacing the gap, that exceed *max* are replaced by *max*, while those who are inferior to *min* are replaced by *min*. This is the last step of the new approach version 2.

A simplistic illustration of this process is provided in figure 5.2. This second part of the new approach is heavily inspired by Sun's strategy [127], with the corresponding differences having already been identified.

Noteworthy to mention that if the last available sample before the gap and the first available sample after the gap have the same value, both versions of this approach would simply consist of a zero-order hold. Therefore, all this process would just be a waste of time and computational resources. Hence, this condition is verified before everything else.

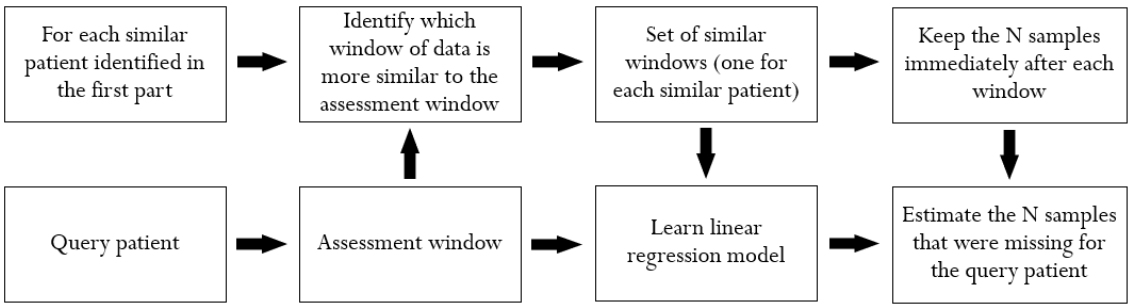


Figure 5.2: Flowchart illustrating the implemented process to estimate the missing samples values. N is the number of missing samples in the gap being handled.

Clustering model development - results and discussion

As pointed out before, for the clustering model development, feature selection procedures were employed. These comprised the assessment of correlations between features and the calculation of mutual information. The correlation assessment's results are presented in figures C.1 to C.6. Given the methodology previously described to consider two features as correlated, only the pair $(Age, Group)$ was deemed correlated.

However, it is interesting to observe that some expected correlations, such as the pairs $(Age, Number\ of\ comorbidities)$ and (Age, ASA), were present in the dataset.

Regarding the calculation of mutual information, the results obtained for a initial clustering model, built considering all 17 features, is displayed in figure 5.3. Given these results, it can be inferred that *Age* contributes more to the clustering than *Group*. Additionally, two sets of "unimportant" features can be distinguished, due to the reduced influence in the clustering, as measured by mutual information. These are, $Set1$ (below 0.01 nat): *diabetes_comorb*, *pulmonal_comorb*, *gastrointestinal_comorb* and *endocrine_comorb*; $Set2$ (below 0.02 nat): *cardiac_comorb*, *thrombotic_comorb* and *neuromuscular_comorb*.

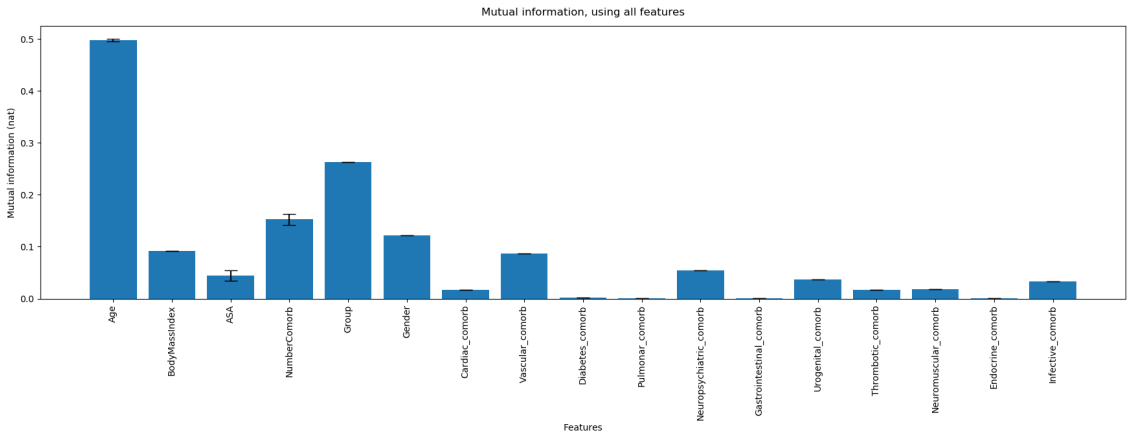


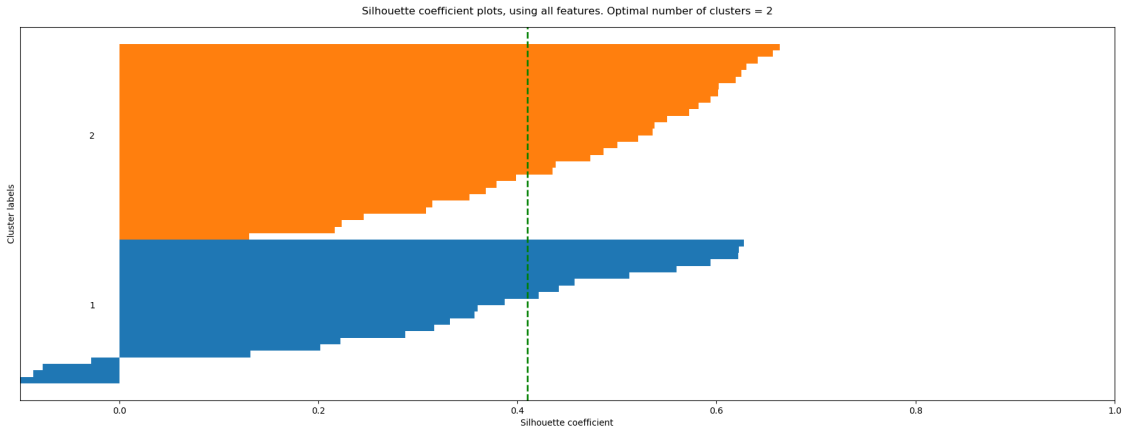
Figure 5.3: Mutual information calculation for a initial clustering model, built considering all 17 features. *NumberComorb* is the number of comorbidities feature. Features that end with *_comorb* are features that refer to the presence or not of the respective type of comorbidity. The error bars represent the 95% confidence interval.

Having these two sets defined, and considering the higher importance of *Age* when comparing with *Group*, clustering models were developed considering the following features: (1) All but *Group*; (2) All but $Set1$; (3) All but $Set1$ and $Set2$; (4) All but *Group* and $Set1$; (5) All but *Group*, $Set1$ and $Set2$.

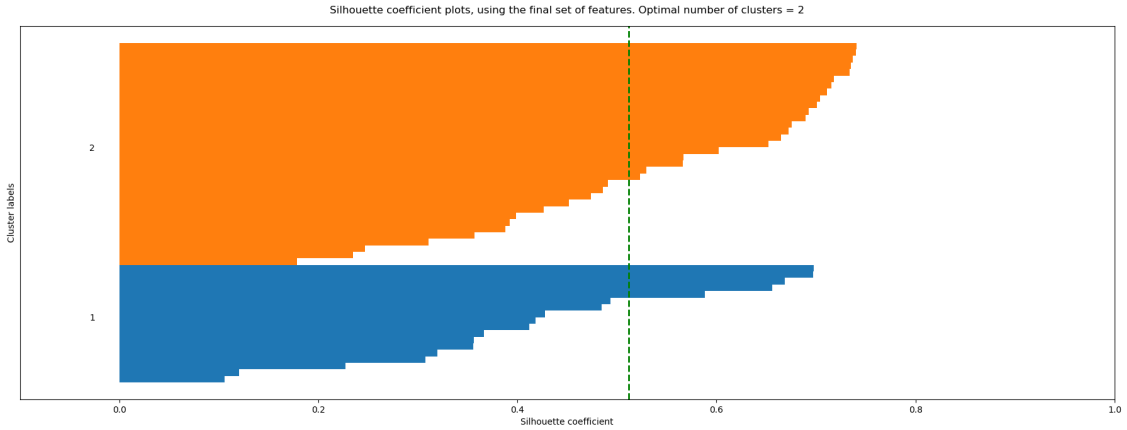
However, since the dataset used was rather small, $Set1$ and $Set2$ could have arisen due to specific characteristics of this population, and might not be representative of what would happen for a larger multicentered cohort. Therefore, clustering models were also developed considering the following features: (6) All but the features that represent the presence or not of certain types of comorbidities (all that end with *_comorb* in figure 5.3); (7) The set of features (6) but *Group*.

The configuration that yielded higher average silhouette coefficient for the respective optimal number of clusters was set as the final clustering model. This corresponded to combination (7). The respective average silhouette coefficient was 0.5 ± 0.2, while the average silhouette coefficient for the initial clustering model considering all 17 features was 0.4 ± 0.2, as shown in figure 5.4. This difference was significantly different ($p - value < 0.01$), which shows that performing the feature selection procedures enhanced the

clustering quality. For both models the optimal number of clusters was 2, as demonstrated in figure 5.5. Features importance in the final clustering model, as measured by mutual information, is displayed in figure 5.6.



(a) Silhouette coefficient for all instances, when clustering was performed including all 17 features. The dashed green line represents the model's average silhouette coefficient, which is equal to 0.4 ± 0.2.



(b) Silhouette coefficient for all instances, when clustering was performed including the features in combination (7) (referred in the figure as final set of features). The dashed green line represents the model's average silhouette coefficient, which is equal to 0.5 ± 0.2.

Figure 5.4: Plot of the silhouette coefficient for all instances in the dataset, both for the initial clustering model (a) and for the final clustering model (b). Both clusterings were performed using the respective optimal number of clusters (see figure 5.5).

As indicated by the positive and relatively large value for the average silhouette coefficient and by the fact that all instances are assigned to the correct cluster (all silhouette coefficients are greater than zero), as evidenced in figure 5.4 (b), it can be stated that the clustering is adequate for the data [133]. Although, an average silhouette coefficient of 0.5 ± 0.2 is still far from the perfect scenario, where it would equal 1, and corresponds to a situation where some points cannot clearly be assigned to the respective cluster [134]. This situation can be explained by the fact that clustering with categorical features is much more challenging, as discussed by Huang [89].

Regarding the optimal number of clusters obtained, it was foreseen that more clusters

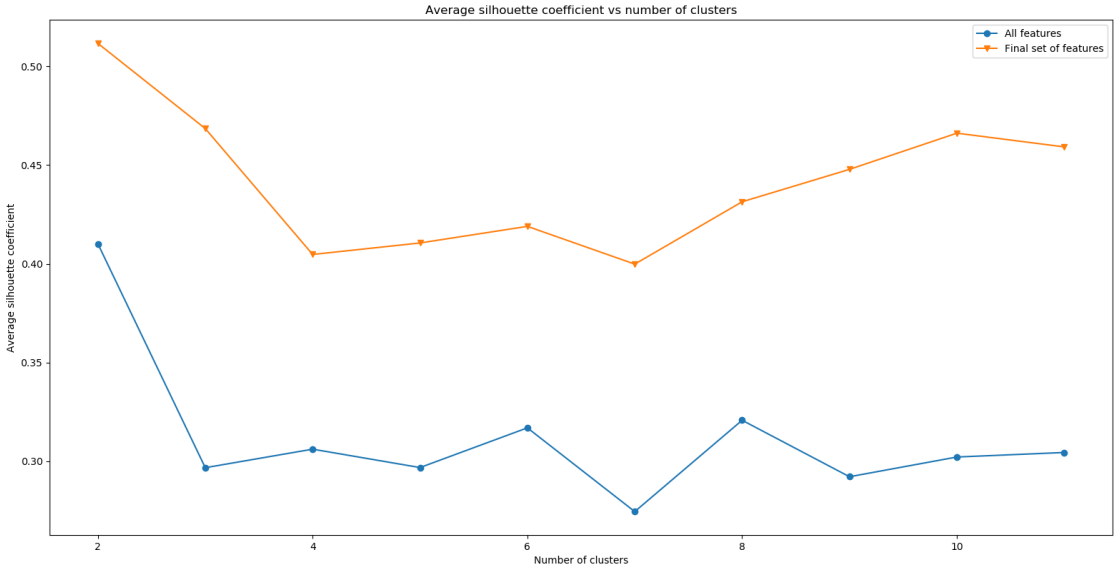


Figure 5.5: Plot of the average silhouette coefficient against the number of clusters. 'All features' refers to the initial clustering model, considering all 17 features. 'Final set of features' refers to the final clustering model, considering the features in combination (7). The optimal number of clusters is 2 for both models.

would be identified, due to the characteristics of the features used, that would permit more "types" of subjects to be found. However, the value obtained might be explained by the small dataset size, which might not contain enough subjects to accurately represent more clusters. This means that with a larger and more representative cohort, the number of optimal clusters would be expected to increase.

With respect to the features importance results, the five relevant features are age, body mass index, ASA, number of comorbidities and gender. By inspecting figure 5.6, it can be concluded that age is, by far, the feature that better explains and contributes for a suitable clustering. Therefore, it's expected that the major demographic difference, when comparing the two clusters populations, occurs on the age feature.

Clusters populations analysis

Table 5.2 presents a comparison of the demographic and contextual characteristics of the two clusters' populations. As expected, the most significant difference (smaller p-value) was found to be on the age feature. Nonetheless, significant differences were also found for the other four relevant features and for the type of surgery (*Group* feature). The significant difference found for *Group* (which was not used in the final clustering model), might be explained by its correlation with age. Since the clustering clearly divided the subjects by age, it would be expected that subjects belonging to the gastroesophageal cancer resection group would also be separated from those belonging to the hip fracture surgery group, since these groups present an average age of 64 ± 10 and 81 ± 7,

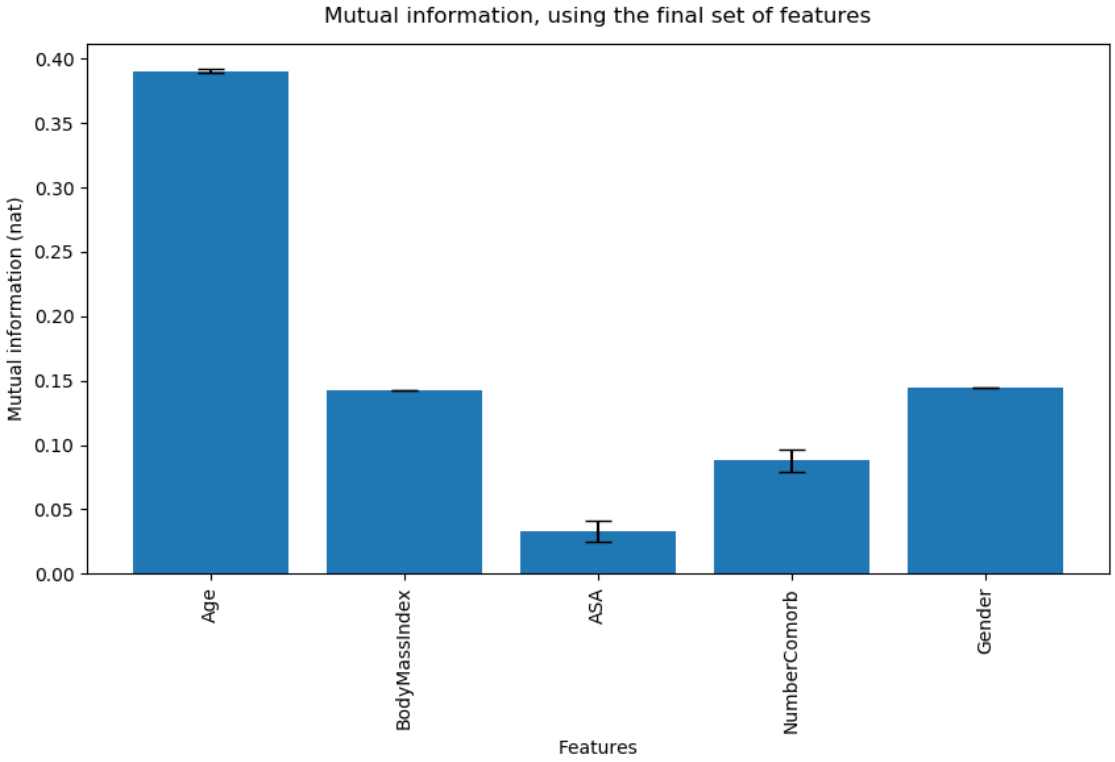


Figure 5.6: Mutual information calculation for the final clustering model, which was built considering the features in combination (7) (referred in the figure as final set of features). $NumberComorb$ is the number of comorbidities feature. The error bars represent the 95% confidence interval.

respectively.

The fact that so many significant differences were found indicates that, demographically and contextually speaking, patients in the same cluster are indeed more likely to be similar to one another than to patients in the other cluster. In the context of the new approach, this is a particularly valuable conclusion, since this novel method is based on the hypothesis that more demographically and contextually similar patients will present more similar vital signs time series [5].

In addition to assessing differences, table 5.2 enables the description of the two clusters populations. Cluster 1 is composed by a population that is younger, more overweight, mostly male, with lower ASA and number of comorbidities, that mainly underwent gastroesophageal cancer resection. Cluster 2 comprises a population that is older, less overweight, mostly female, with higher ASA and number of comorbidities, that mainly underwent hip fracture surgery. These descriptions can be summarized by providing a conceptual description of each cluster type subject, based on the clusters centroids [88]: **Cluster 1 type subject** - 57.7 years, 29.6 kg/m^2, ASA of 2.3[3], 2.2 comorbidities and male gender; **Cluster 2 type subject** - 77.2 years, 23.5 kg/m^2, ASA of 2.7, 3.6 comorbidities

[3] ASA of 2 represents a patient with mild systemic disease, while an ASA of 3 represents a patient with severe systemic disease [135]

and female gender.

Table 5.2: Comparison between patients in the identified clusters. This comparison involves the five features included in the final clustering model, the type of surgery and the subject situation ("Event" or "Non-Event"). The existence of significant differences between the two groups was assessed.

	Cluster 1	Cluster 2	p-value[a]
Number of subjects	18	34	—
Age, years (mean ± SD)	58 ± 9	77 ± 8	< 0.01
Body mass index, kg/m^2 (mean ± SD)	30 ± 6	24 ± 4	< 0.01
ASA, median (first quartile / third quartile)	2.0 (2.0/2.8)	3.0 (2.0/3.0)	< 0.05
Number of comorbidities, median (first quartile / third quartile)	2.0 (1.0/3.0)	3.5 (3.0/4.8)	< 0.01
Gender, male	16 (88.9%)	12 (35.3%)	< 0.01
Type of surgery, gastroesophageal cancer resection	17 (94.4%)	16 (47.1%)	< 0.01
Subject situation, "Event"	5 (27.8%)	10 (29.4%)	0.91[*]

SD - standard deviation, ASA - American Society of Anesthesiologists class.
[a] significance assessed using the Wilcoxon rank sum test at 5% significance level.
[*] not significant.

5.1.2.2 Technique selection - error study

Methods

Given that no method can be considered the absolute best across all possible datasets, six of the already mentioned strategies were implemented: last value, linear interpolation, median over the previous 1-hour, average over the previous 1-hour and the two versions of the new approach. The 1-hour value was based on prior research [127].

These strategies needed now to be tested, to evaluate which one(s) should indeed be used. Additionally, it was also necessary to define a maximum gap duration to handle, which means missing data periods longer than that would remain unprocessed. Both these assessments were done by performing an **error study**. This is, 200-minutes long segments of data without missing samples were extracted from all subjects' four vital signs time

series. Then, the following procedure was completed for each of those segments:

1. a missing data period was simulated and handled with one of the implemented strategies.
2. the relative error rate (equation 5.4) was calculated.
3. step 1. and 2. were repeated several times with a different random location for the simulated missing data period.
4. finally, all relative error rates obtained by step 2. were averaged to obtain the average relative error rate for the segment, e_{seg}.

$$e = \sum \frac{(x_i - \hat{x_i})^2}{x_i^2} \tag{5.4}$$

where e is the relative error rate, x_i is the true value of sample i and $\hat{x_i}$ is the estimated value of sample i using one of the strategies. The summation applies to all samples belonging to the simulated missing data period.

This process was executed for the six different strategies and for gap durations of 5, 10, 15, 20, 25, 30 and 60 minutes. The final relative error rate for each combination of strategy and gap duration was, then, obtained by averaging the set of e_{seg} values returned by the above procedure, using that strategy and that gap duration. The respective 95% confidence intervals were calculated.

Additionally, the exact same procedure was performed but considering only segments from one vital sign at a time, since there was the possibility that each vital sign was linked to a different most suitable strategy.

A final note on this procedure can be made to justify the decision on the value of 200 minutes for the segments duration. The largest gap duration tested was 60 minutes and some strategies (average and median over the last 1-hour, and the two versions of the new approach) require a previous portion of 60 minutes of available data. This means that a segment of, at least, 120 minutes was necessary. A slightly higher value was chosen so that the random missing periods could always have some variability. Nonetheless, any value greater or equal to 120 would suffice.

Results and discussion

Figure 5.7 shows the results of the error study just described. It is observed that linear interpolation is the best technique across all gap durations tested, in terms of relative error rate. Also, after linear interpolation, new approach version 2 was the method that presented lower relative error rate. Despite this not directly meaning that these are the two methods that will produce the best performance in the deterioration prediction task, it indicates that these methods more accurately correct the vital signs time series, which intuitively makes them the preferred choice in this context.

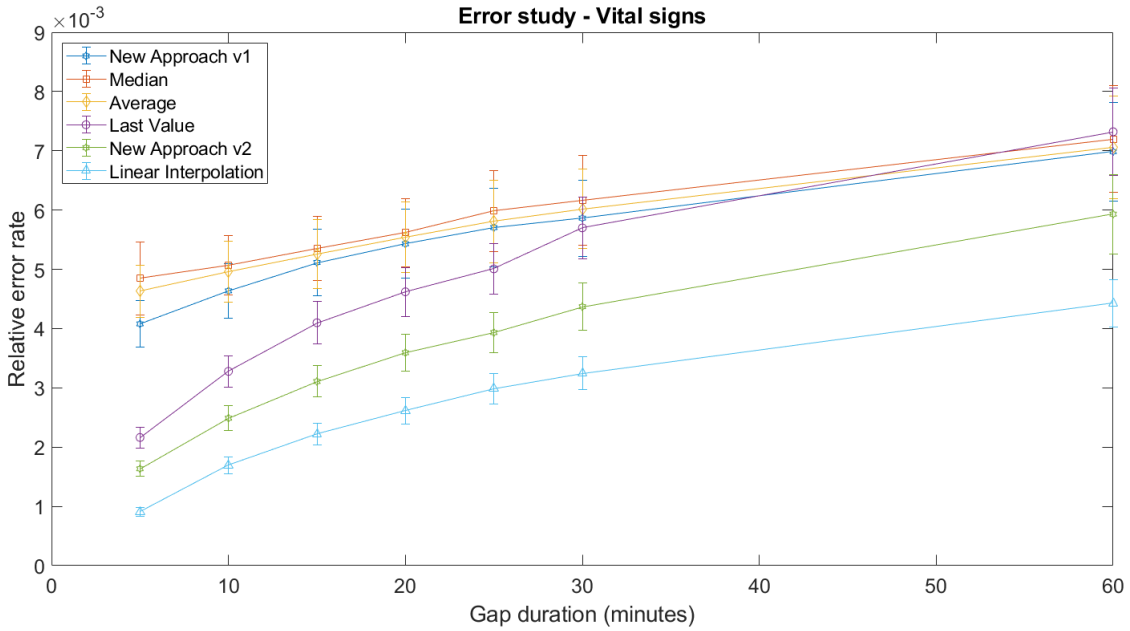


Figure 5.7: Results of the error study performed for the selection of an adequate technique to handle missing data periods in the vital signs time series. This study's methodology and the six techniques being tested are described in 5.1.2.2. The error bars represent the 95% confidence interval.

Additionally, and as illustrated in figures C.7 to C.9, it was hypothesized that the new approach captures the signals' nature and variability to a greater degree than linear interpolation. Hence, it could improve performance in the deterioration prediction task by providing better representations of the missing data periods.

Therefore, and in order to assess which one yielded superior prediction performance, the two versions of the new approach and linear interpolation were the techniques actually implemented. Hence, from now on, three different datasets are distinguished:

- Linear Interpolation dataset (LinInt) - where all patient's vital signs were preprocessed so that gaps $\leq$ 60 minutes are corrected using linear interpolation.
- New approach version 1 dataset (NApp1) - where all patient's vital signs were preprocessed so that gaps $\leq$ 60 minutes are corrected using (1) the new approach version 1, if the 60 minutes segment before the gap doesn't contain missing values; (2) linear interpolation, otherwise.
- New approach version 2 dataset (NApp2) - where all patient's vital signs were preprocessed so that gaps $\leq$ 60 minutes are corrected using (1) the new approach version 2, if the 60 minutes segment before the gap doesn't contain missing values; (2) linear interpolation, otherwise.

The same strategies were employed for the four vital signs, since the techniques tested presented similar relative results across the four vital signs (see figures C.10 to C.13).

The error study results obtained for HR and SpO_2 (see figures C.10 and C.13) can be compared with Sun's [127] method, since these two vital signs were used in their study. They adopted the same formula for the error calculation (equation 5.4) and reported error results for a 10% missing rate, which corresponds to the 20 minutes gap duration in this study. Their data was also sampled at 1-minute intervals. They achieved a relative error rate of 0.0014 and 0.0008 (standard deviations not reported) for HR and SpO_2, respectively, while the results of the new approach version 2 were 0.0025 and 0.0002, respectively. These are very similar results, which can be justified by the similarities between the second part of the two strategies. However, the new approach developed in this thesis has the advantage of neither requiring labels provided by experts nor requiring the patient to already have available data, for the identification of a set of similar patients. Instead, it employs an unsupervised learning-based clustering approach in combination with demographic and contextual features, which are promptly acquired at patient admission.

Besides the two above-mentioned strategies, other personalized approaches to deal with periods of missing data, in vital signs, have been developed. In fact, the personalized Gaussian processes-based framework developed by Clifton et al. [98] also achieved a smaller error than simpler approaches, like imputing the patient average. Sow et al. [125] employed a forecasting method based on fading memory polynomial filters, which also proved capable of making accurate sample estimations, in physiological data, for gaps as large as 60 minutes.

All these studies demonstrate the added value of personalized methods for correction of missing data periods in vital signs time series. This gains relevance when the discussion concerns surgical patients, due to their particular characteristics [5], [97]. Thereby, in this context, the adoption of such personalized methods is suggested.

Nonetheless, in studies where these are not to be employed for some reason, a suggestion can be provided for the use of linear interpolation. The reason for this is related to the fact that in the reviewed literature, where continuous monitoring of vital signs was carried out, other simple approaches were preferred (median [101], last value [102] and average [97]). However, as demonstrated in figure 5.7, linear interpolation is a more accurate technique for the estimation of missing samples in the vital signs time series.

5.2 QRS complex amplitude

The QRSa signal is not used very often, especially for the early detection of complications (only one study [136] was found and with a very particular and distinct goal). In addition to that, the preprocessing focus is usually on the ECG signal, from which the QRSa is then extracted, not on the QRSa signal itself. Therefore, no preprocessing strategy for this time series was found.

However, due to its unsatisfactory quality in this dataset, additional preprocessing had to be implemented. Since no guidance on this procedure was available, a strategy

identical to the one applied for the vital signs was implemented: **artifact removal**, to simply remove the occasionally appearing outliers, and **missing data handling**, once again, to deal with periods of absence of data.

After these two preprocessing stages, the signal still undergoes the process of **normalization**, which is explained further ahead.

5.2.1 Artifact removal

As for the vital signs, the first step was to remove obvious outliers. Given the diverse physiological ranges this measure can have across individuals, this was performed using very conservative thresholds: if any sample was below 0 mV or above 6 mV, it would be excluded and replaced as missing value.

Next, a variety of moving mean filters and median filters were tested, to try and mitigate the presence of high frequency noise and possible outliers still remaining. In the end, a 20 samples-based moving mean filter was applied.

5.2.2 Handling missing data

Since no information was available on how to handle missing data periods in this signal, the same strategies as for the vital signs were tested: imputing the last value before the gap (or zero-order hold), linear interpolation and imputing the median or average over previous values. The new approach couldn't be applied here because this is an unevenly sampled time series and, currently, the new approach only works for evenly sampled signals.

5.2.2.1 Technique selection - Methods

The same procedure to assess which method should be implemented was performed, i.e., an **error study**. The only differences between this study and the one performed for the vital signs are (1) the gap durations tested were 1, 2, 5, 10 and 30 minutes, since the sampling frequency of this signal is much higher; (2) the average and median strategy were applied to the previous 2 minutes of data, instead of the previous 1 hour; (3) instead of 200-minutes long segments of data, 80 minutes segments were extracted, since any value greater or equal than 32 minutes would suffice (2 minutes of data required by the average and median strategy, plus the largest gap duration tried, 30 minutes). For the same reason as before, a slightly higher value was chosen.

The value of 2 minutes to average or median over was itself selected based on a similar error study, where different past interval durations were tested. In particular, 1, 2, 5 and 10 minutes to average or median over were experimented. This was the solution found

to select an appropriate value, since unlike with the vital signs, no previously reported value for the interval duration to average/median over was available.

5.2.2.2 Technique selection - Results and discussion

The error study results are presented in figure 5.8. The average and median techniques plotted there were applied considering the 2 minutes of data before the gap, due to the results displayed in figures C.14 and C.15. In those, the confidence intervals show that it is practically irrelevant which interval of data is considered, at least between the 1-minute and 2-minutes intervals and for large gaps. Nonetheless, the 2-minutes interval was selected as it presents slightly better results for larger gaps.

With a more profound inspection of figure 5.8, it can be understood that linear interpolation was the employed strategy, as it demonstrates the lowest relative error rate across all gap durations tested. The maximum gap duration to handle was decided to be 10 minutes (600 seconds in figure 5.8), since this is the last tested gap duration that presented a relative error rate lower than 0.01 (which corresponds to 10% average error, since equation 5.4 is squared).

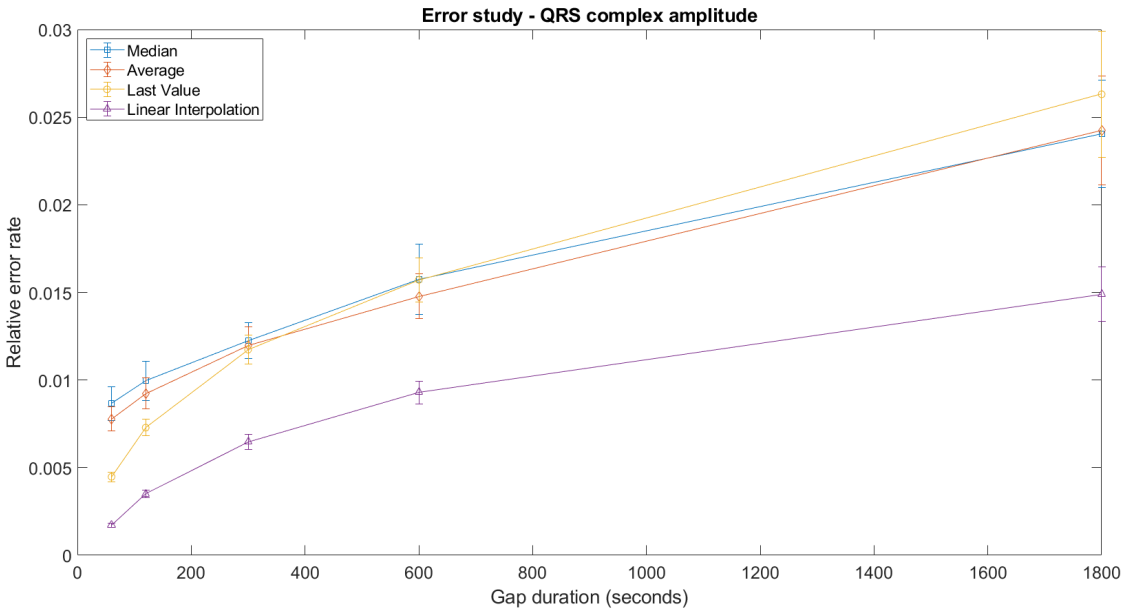


Figure 5.8: Results of the error study performed for the selection of an adequate technique to handle missing data periods in the QRSa time series. This study's methodology and the four techniques being tested are described in 5.2.2. The error bars represent the 95% confidence interval.

Once again, the fact that linear interpolation is the technique with the lowest relative error rate does not imply that it is the one that yields a higher performance in the

prediction task. However, it corrects the QRSa time series more properly, making it the preferable option.

Since no other study was found where this time series was preprocessed, no comparison discussion can be performed. Still, the use of linear interpolation, for the correction of this time series, can be suggested for future studies that intend to work with it.

5.2.3 Normalization

The process of signal normalization, also known as Min-Max scaling[4], ensures all the signal's samples are within a fixed range ([0,1]), despite being highly affected by outliers. This procedure was only applied after each subject's signal was divided in windows (explanation in section 6.1).

The addition of this extra preprocessing stage was prompted by the fact that this measure's physiological range might be very distinct from subject to subject, as mentioned in subsection 2.3.1.

5.3 RR interval (RRI)

The strategy to preprocess the RRI time series generally comprises three stages: **ectopic beats/artifact removal**, to exclude false beats and unreliable data, **missing data handling**, again, to deal with periods of absence of data, and **detrending**, to remove any non-stationarity present in the signal, which can affect power spectrum estimations [70]. Some studies [67], [68] mention an additional fourth stage, resampling, due to the RRI time series being unevenly sampled. This is sometimes done for the extraction of frequency domain features, since the techniques commonly employed for power spectrum estimation, like the fast Fourier Transform, require evenly sampled time series [137]. However, there are reports [137] that resampling and the following use of conventional spectral methods introduces significant errors in the RRI power spectrum estimation. Therefore, this stage was skipped and a more adequate spectral estimation technique, that supports unevenly sampled data, was utilized for the extraction of frequency domain features, the Lomb–Scargle periodogram. This technique was shown to provide better power spectrum estimations [137].

Regarding the first preprocessing stage, this signal can be affected by two types of artifacts, which can originate from physiological or technical reasons [70]. The latter are usually the result of sensor issues or patient movement. The former most often arise from missed beats or ectopic beats.

Ectopic beats are a consequence of disturbed electrical activity in the heart. Usually, the cells in the sinoatrial node are the ones who initiate the propagation of the electrical

[4] applying to all samples: $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$, where x_{norm} is the sample value after rescaling, x is the sample value before rescaling, x_{min} is the minimum value in the set of samples being considered and x_{max} is the maximum value in the set of samples being considered

impulses that will result in a heartbeat. However, if this process is started in the ventricles or in the atria, a premature ectopic beat might occur. These are not necessarily related with pathological causes but their presence is a major source of errors when measuring heart rate variability using the RRI signal. These ectopic beats are easily identified in the RRI time series, since these are represented by a short duration interval, followed by a long duration interval (compensatory pause), before the return to the previous baseline rhythm [70]. This can be observed in figure C.16, where several ectopic beats are shown. Not excluding these ectopic beats and the other artifacts from the RRI time series analysis can compromise the reliability of the features that are extracted from this signal [70], [138].

5.3.1 Ectopic beats/artifact removal

As for the other types of data used in this thesis, the first step for artifact removal was the exclusion of obvious outliers. This was accomplished by excluding and replacing as missing value, every sample that either was below 300 ms or above 2000 ms. These values were chosen based on the thresholds applied for HR.

The next step was the correction of false beats, which mostly originate from premature ectopic beats. This was done through the application of a novel technique, developed during this study.

5.3.1.1 Novel technique

Methods

This technique combines a selective median filter with a previously developed impulse rejection filter, where an additional threshold modification was performed. The method's novelties are, then, the application of the selective median filter before applying the impulse rejection filter, and the threshold modification.

The selective median filter consists of, first, generating a RRI time series filtered with a 10 samples-based median filter, $x_{med}(n)$, and then replacing every sample in the original RRI time series, $x(n)$, according to:

$$\hat{x}(n) = \begin{cases} x_{med}(n), & if \frac{|x_{med}(n)-x(n)|}{x(n)} > 0.2 \\ x(n), & otherwise \end{cases} \tag{5.5}$$

where $\hat{x}(n)$ is the obtained time series filtered with the selective median filter.

Then, the impulse rejection filter is applied to $\hat{x}(n)$. This filter had previously been designed [139] and used in other studies [140]. In practical terms, the filter is applied iteratively in 5 minutes segments, until the entire RRI time series is preprocessed. It consists of, first, calculating the following test statistic:

$$D(n) = \frac{|s(n) - s_m|}{1.483 \times med\{|s(n) - s_m|\}} \tag{5.6}$$

where $s(n)$ is a 5 minutes segment from the RRI time series, med$\{\cdot\}$ is the median operator and $s_m = med\{s(n)\}$.

Then, the filtered segment, $\hat{s}(n)$, is calculated as:

$$\hat{s}(n) = \begin{cases} s(n), & D(n) < \tau \\ m(n), & D(n) \geq \tau \end{cases} \tag{5.7}$$

where τ is a specified threshold. This is the threshold that was modified. It was reported the use of $\tau = 4$ but $\tau = 2$ was used instead, because, visually, it seemed to consistently yield better filtering for the data used in this thesis. $m(n)$ is calculated as:

$$m(n) = med\left\{s(n+m) : |m| \leq \frac{w_m - 1}{2}\right\} \tag{5.8}$$

where w_m is the length of the window centered around n where the median operator is applied. $w_m = 5$ was used, as recommended.

For a more detailed explanation on this filter see McNames et al. [139].

Results and discussion

The results of (1) applying only the selective median filter; (2) applying only the impulse rejection filter; (3) applying the entire novel technique, were only assessed qualitatively, by visual inspection.

A comparison example between the three methods and the original signal is provided in figure 5.9. For a better separate comparison between the novel technique result and each of the other time series, in the given example, consult figures C.16 to C.18.

The provided example in figure 5.9 illustrates what was observed for the generality of the RRI signals analyzed in this thesis. This is, visually, the novel technique seems to attain a superior rectification of ectopic beats and artifacts than the two filters alone. However, that cannot be affirmed, since no quantitative validation was performed, mostly due to time constrictions.

A future quantitative validation of this technique might be achieved, for example, by comparing the percentage of ectopic beats corrected by each technique, after these being marked manually or through automatic algorithms, such as the Pan & Thompkins algorithm[5] [141]. Other solution, more driven to the usage of the RRI signal for the extraction of features in the context of ML models development, might pass by the artificial introduction of these artifacts into a clean time series. Then, features can be extracted from the original clean time series and from time series preprocessed with each of the above-discussed techniques, and differences between the features values can be assessed.

[5] usually employed for the detection of R-waves in the ECG signal

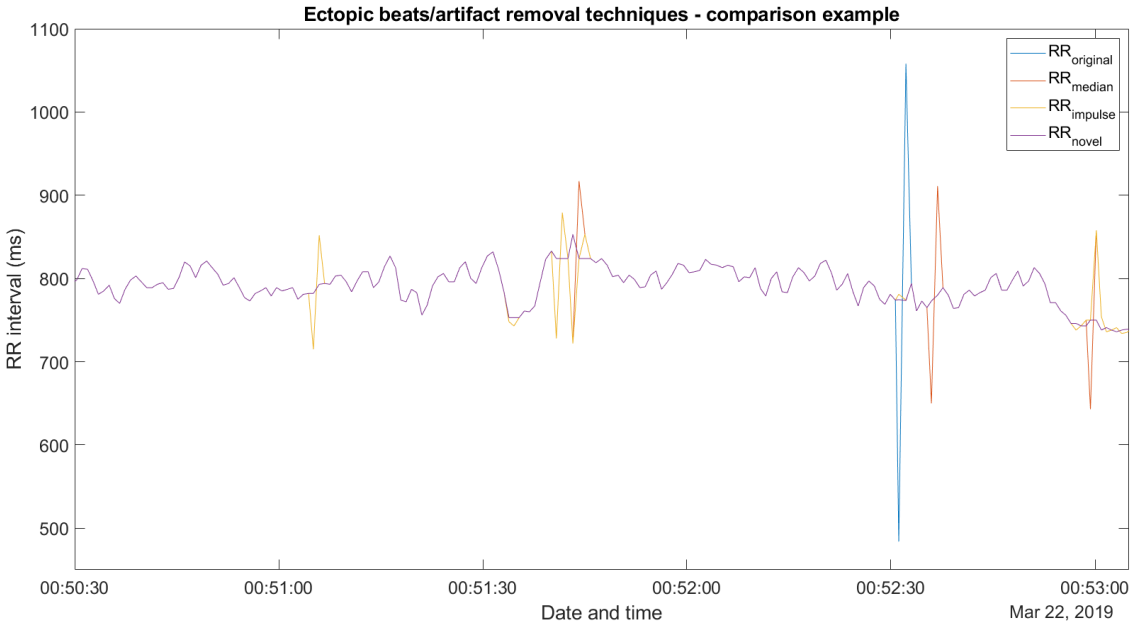


Figure 5.9: Comparison example between an unprocessed RRI time series ($RR_{original}$) and the same time series preprocessed with only the selective median filter (RR_{median}), only the impulse rejection filter ($RR_{impulse}$) and the novel technique (RR_{novel}), which combines both filters.

The filter in which the novel technique depends, is not the only validated strategy that can be found in the literature for ectopic beats/artifact removal. As a matter of fact, Logier et al. [142] developed a filter that replaced up to 90% of erroneous beats. Although, it was tested in a rather small dataset. Additionally, the strategy of excluding beats that differ by more than 20% from adjacent beats was already validated in RRI data from rodents [138] and employed in studies involving humans [67]. Despite these being already validated approaches, the development of novel methods, such as the one discussed in this thesis, might bring advances to the field of physiological time series preprocessing.

5.3.2 Handling missing data

The most common approach to deal with periods of missing data in the RRI time series is the use of interpolation methods. These have previously been recommended [70], especially if a frequency domain analysis will be performed later, which is the case.

5.3.2.1 Technique selection - Methods

That being said, linear interpolation was tested. Once again, an **error study** was performed, but since now there was only one strategy being tested, the study only meant to define the maximum gap duration to handle. 5, 15, 25, 35, 45 and 60 seconds gap durations were experimented, since correcting long periods can modify the signal's frequency content [142]. Additionally, this time, 15-minutes long segments of data were

extracted. Since any segment greater or equal to around 62 seconds would suffice (for the interpolation, one sample before the gap and one after are required, plus the largest gap duration tried, 60 seconds), a slightly higher value was chosen, for the same reason as before.

5.3.2.2 Technique selection - Results and discussion

The error study's results are displayed in figure 5.10. Considering the relative error rate values obtained, gaps could be handled until the 60 seconds duration. However, even for this short gap duration, the risk of modifying the signal's frequency content could be incurred. Hence, only missing periods ≤ 15 seconds were handled, as in prior research [142].

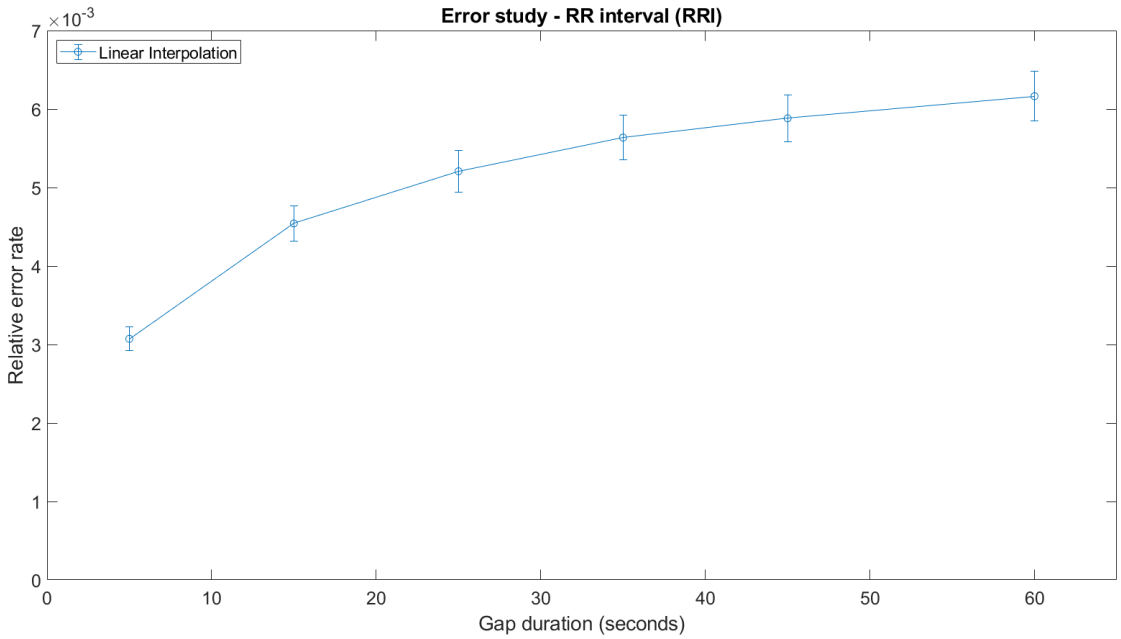


Figure 5.10: Results of the error study performed for the selection of an adequate maximum gap duration to handle missing data periods in the RRI time series. This study's methodology is described in 5.3.2.1. The error bars represent the 95% confidence interval.

Besides linear interpolation, other interpolation methods, such as cubic spline interpolation, can be applied for the correction of missing data periods in the RRI signal [70], [143]. In fact, Morelli et al. [143] performed a very complete comparison of the effects of different types of interpolations on the corrected RRI time series and respective features estimations. Linear and quadratic interpolation have shown to be the techniques that induce less errors in the RRI time series and in the features estimations, respectively.

5.3.3 Detrending

This preprocessing stage was applied to obtain a detrended RRI time series, which was used only to extract the frequency domain features. This must be performed because the

signal's non-stationarities can affect the power spectrum estimation [70].

The method used for detrending was the wavelet detrending method. The Daubechies-6 wavelet with four levels of decomposition was used, as in prior research [67]. This method decomposes the signal into approximation and detail coefficients, using the discrete wavelet transform. Each decomposed sub-band has an associated set of frequencies, where the highest level of approximation coefficients represent the lowest frequencies. Those coefficients are then set to zero, and the inverse discrete wavelet transform is applied to reconstruct the signal [67], [68]. However, as the lowest frequency coefficients were eliminated, the reconstructed time series will have its baseline trend removed.

5.4 Preprocessing summary

Figures 5.11, 5.12 and 5.13 are simplistic illustrations that intend to summarize the preprocessing stages each type of data went through.

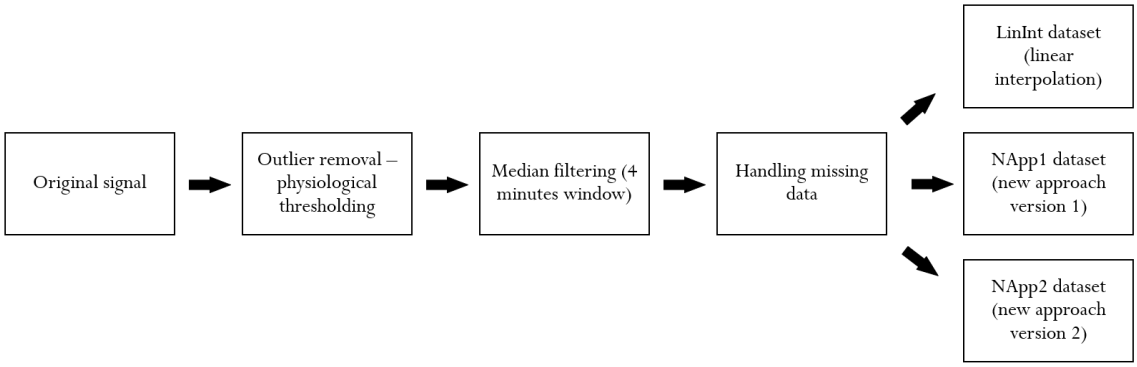


Figure 5.11: Flowchart illustrating the implemented procedure to preprocess the vital signs time series.

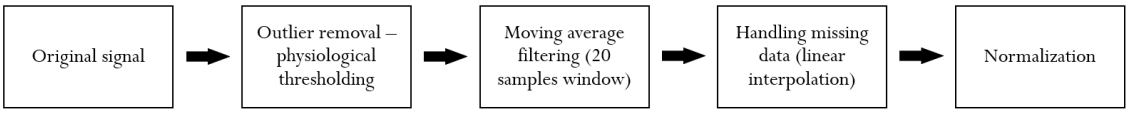


Figure 5.12: Flowchart illustrating the implemented procedure to preprocess the QRS complex amplitude (QRSa) time series.

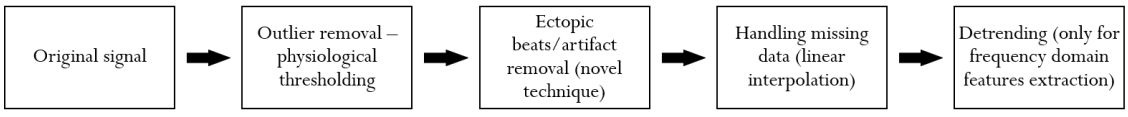


Figure 5.13: Flowchart illustrating the implemented procedure to preprocess the RR interval (RRI) time series.

Chapter 6

Warning system development

The development phase of the warning system for clinical deterioration detection, created during this thesis, depended on the completion of four main stages. The first two, **data acquisition** and **preprocessing**, were already described in chapters 4 and 5, respectively.

This chapter starts by clarifying the prediction strategy. Next, it details the methodology and discusses the results of the remaining two stages. These are the **features extraction** procedure and the **development of a ML-based prediction model**. The best prediction models developed are characterized and the respective results are compared against reviewed work in the field. Then, the assemble of the final warning system is discussed and an estimation of the warning system's possible usage frequency is provided. Finally, the study limitations and suggestions for a future work are identified.

6.1 Prediction strategy

As discussed in subsection 2.4.1, for the development of a prediction model, the dataset D has to be generated. Hence, the patients' continuous time series have to be transformed in a set of observations, which are represented by a group of relevant data properties, the features. One way of doing this, would be to have an observation for each patient or to consider a new observation every time a new set of measurements was available. However, both these strategies have drawbacks. The latter might result in an unbearably large number of observations. The former requires all patient's data to be analyzed as a whole, which might be misleading and hinder the prediction task, since in the same hospital stay a patient can both have stable and deteriorating periods.

Considering these drawbacks, the implemented prediction strategy was an intermediate between those two strategies. It involved splitting the patients data in windows, using a sliding-window approach, with 1 hour steps and 12 hours window size. This

means that multiple windows are extracted from the same patient, hence every patient is contributing with various observations for the dataset D (each window is an observation). For the "Event" subjects only the data until the deterioration event was considered, while for the "Non-Event" subjects all available data was used.

The resulting set of windows was then labeled using a discrete time analysis. This is an approach that has been used before in this context [78], [103], [104], [108], [109] and its concept consists of analyzing data from a past interval, e.g., previous 12 hours, to try and predict if a deterioration event will occur in a certain future interval, e.g., the following 12 hours. Applied here, this approach meant labeling each window with 1 (positive class), if a deterioration event occurred anywhere in the 12 hours interval following the end of the window, t_0. Otherwise, the window would be labeled 0 (negative class). Figure 6.1 illustrates this process. This strategy implies that deterioration is being predicted with 12 hours or less in advance and that "Event" subjects also contribute with 0-labeled windows (those where the deterioration event was more than 12 hours away from t_0). Additionally, if applied in real time, this prediction strategy would mean analyzing only the most recent 12 hours of patient's data to predict if a deterioration event would occur in the following 12 hours. Both these values selection was guided by values used in prior research (see table A.1).

In summary, this strategy's outcome is a set of windows, where each of them is already coupled with an output variable, y_i. Given that this is a binary classification problem, the values of y_i were restricted to be labels (0 or 1). The number of windows obtained after this procedure is presented in table 6.1.

This initial set of windows, however, could contain windows where some of the six physiological signals being considered (HR, RR, BTemp, SpO_2, QRSa and RRI) were missing for long periods. Therefore, the following process was completed for every window:

1. apply the respective physiological thresholding to the six time series.

2. apply the following **acceptance criterion**: if all six time series have at least 50% of available samples in the window (not missing value), accept the window. Otherwise, exclude it.

The set of windows that passed the acceptance criterion would then go through the remaining preprocessing stages, described in the previous chapter. After that, these windows were ready to enter the features extraction stage. The number of windows obtained after this procedure is also presented in table 6.1.

6.2 Features extraction

Having a set of windows/observations identified, the only thing that was missing to have the dataset D ready to be used for model development was extracting features that could properly represent each window.

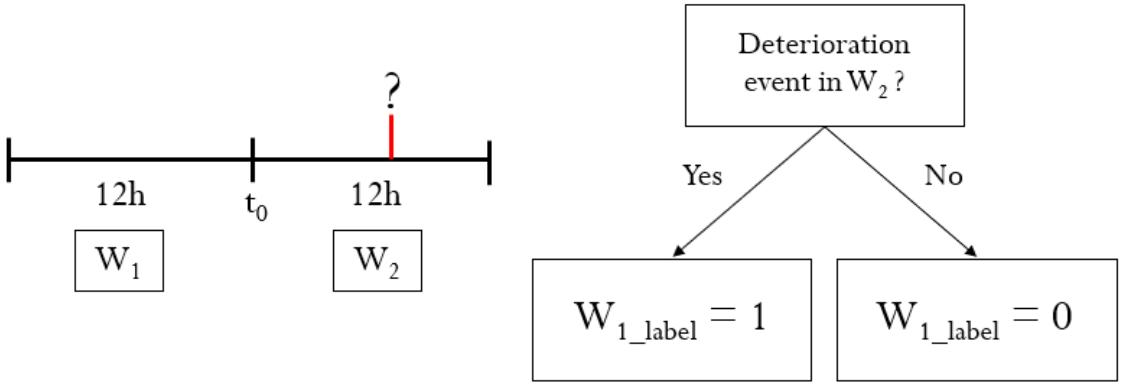


Figure 6.1: Illustration of the windows labeling process. This illustration regards the labeling of an arbitrary window, W_1, where t_0 represents its last timestamp. The question mark intends to emphasize that this process involves assessing if a deterioration event occurred in W_2, which is the following 12 hours interval. W_{1_label} is the label attributed to window W_1.

The methodology to attain the best set of features consisted in programming and extracting a vast amount of them, and then applying feature selection procedures. These feature selection procedures are model specific, so they are discussed further ahead.

In total, 104 features were implemented. These comprise both numerical and categorical features, which are detailed and described in appendix B.

When extracting these features from the set of windows identified, it sometimes happened that some feature couldn't be successfully extracted. This could occur, for example, when extracting a feature that required the most recent 15 minutes of data for a window that had no available data in that period. Consequently, the windows where at least one feature couldn't be extracted had to be excluded. The number of windows obtained after these exclusions is presented in table 6.1. This corresponds to the final set of windows utilized for the development of the prediction models. This set presents a classes ratio of around 1:50, i.e., one 1-labeled window for fifty 0-labeled windows.

In summary, the dataset D includes 2152 observations, each represented by an appropriate feature vector, $\boldsymbol{x_i}$, and already coupled with an output label, y_i. Each feature vector has 431 elements/dimensions, from which 418 are numerical and 13 are categorical. These 431 elements are the result of the extraction of the 104 features.

Table 6.1: Number of available windows for model development. The final set of windows used corresponds to the last row of the table.

	Number of available windows	**Number of available windows labeled 1**
Initial set	5876	156
After acceptance criterion	2405	48
After features extraction	2152	45

6.3 Datasets preparation and variations

Despite the discussion made so far was being made referencing a dataset D, three different datasets can be distinguished, as mentioned in 5.1.2.2: LinInt, NApp1 and NApp2. They all have the same number of observations and the same number of features dimensions. What might differ between them are the values themselves of the features extracted from the vital signs.

That being said, these three datasets went through four preparation steps:

- **relevance check** - all dimensions with variance zero across the whole dataset were excluded. The reason for this is that these specific dimensions do not contain any information that can help distinguish between the two classes.
- **dataset partition** - the dataset was split into 70% training set and 30% test set. The classes ratio was kept similar in both sets (1:50).
- **numerical features rescaling** - the numerical features were rescaled, through z-score standardization[1]. Only the training set observations were used to calculate the standardization parameters.
- **categorical features encoding** - for LR models, the categorical features were transformed into dummy variables via one-hot encoding[2].

In addition to that, correlations between features were assessed. The exact same procedure was employed here, as for the assessment of correlated features in the clustering model development (see 5.1.2.1), with two exceptions. First, the correlation threshold to deem two features as correlated was set to 0.9. Second, since all categorical features could be interpreted as ordinal, the Spearman rank correlation coefficient and the Kendall rank correlation coefficient tau-b (both range from -1 to 1) were implemented for the numerical-ordinal and ordinal-ordinal correlations assessment.

Since the datasets would still undergo feature selection procedures, the information obtained from the correlation assessment was only used for *a posteriori* analysis on the features that remained.

Datasets variations

As mentioned before, the three datasets present a classes ratio of 1:50, where the minority class is the deterioration cases (positive class). This poses severely imbalanced datasets, which can hamper the prediction task, especially the prediction of the positive

[1] $z = \frac{x-\mu}{\sigma}$, where x is the feature value before rescaling, z is the feature value after rescaling, μ is the feature mean value in the dataset and σ is the feature standard deviation in the dataset.

[2] each categorical feature is transformed into $k-1$ dimensions, where k is the number of different categories for the respective feature. Each new dimension either contains 0 or 1, to indicate the presence or absence of that category in the observation being considered.

class [144], [145]. As can be verified by inspecting appendix A, this is a common situation in this type of studies, which suggests this is an intrinsic and naturally occurring problem in these datasets [144], [145].

To address this issue, two solutions were tested:

- **undersampling** of the majority class, which combines the minority class with only a subset of the majority class [108] (see table 6.2). This was performed to obtain three additional datasets variations, with classes ratios of 1:1, 1:4 and 1:10.

- implementing **boosting techniques**, since its use had been highly recommended for imbalanced datasets problems [79], [144], [145].

Table 6.2: Number of windows used for model development, depending on the classes ratio being considered.

Ratio	Number of windows labeled 0	Number of windows labeled 1	Total number of windows
1:50 (original)	2107	45	2152
1:10	450	45	495
1:4	180	45	225
1:1	45	45	90

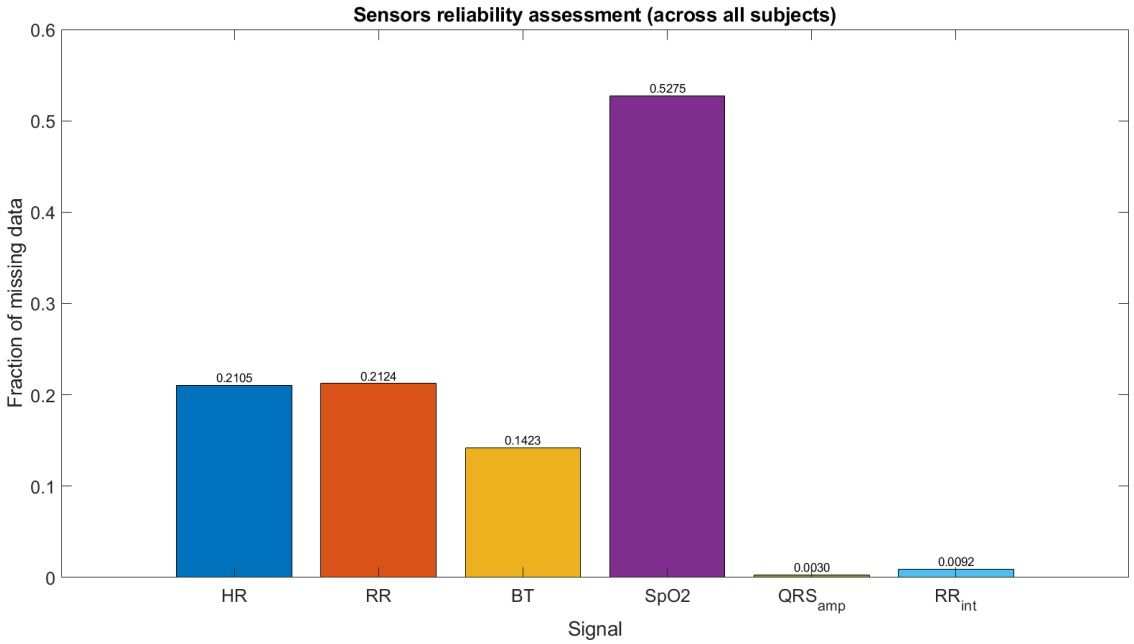
Additionally, one more dataset variation was produced. This was the result of a reliability assessment on the sensors utilized, based on the fraction of the acquired data that was missing. Measures outside the physiological thresholds already discussed were considered missing values.

This assessment was prompted by the fact that it was visible that some signals, especially SpO_2, were missing for very long periods and across many subjects. The assessment's results are displayed in figure 6.2. Figure 6.2 (a) shows the fraction of missing data points across all subjects (per signal), while figure 6.2 (b) shows the mean and respective 95% confidence interval of the fraction of missing data points per subject (per signal).

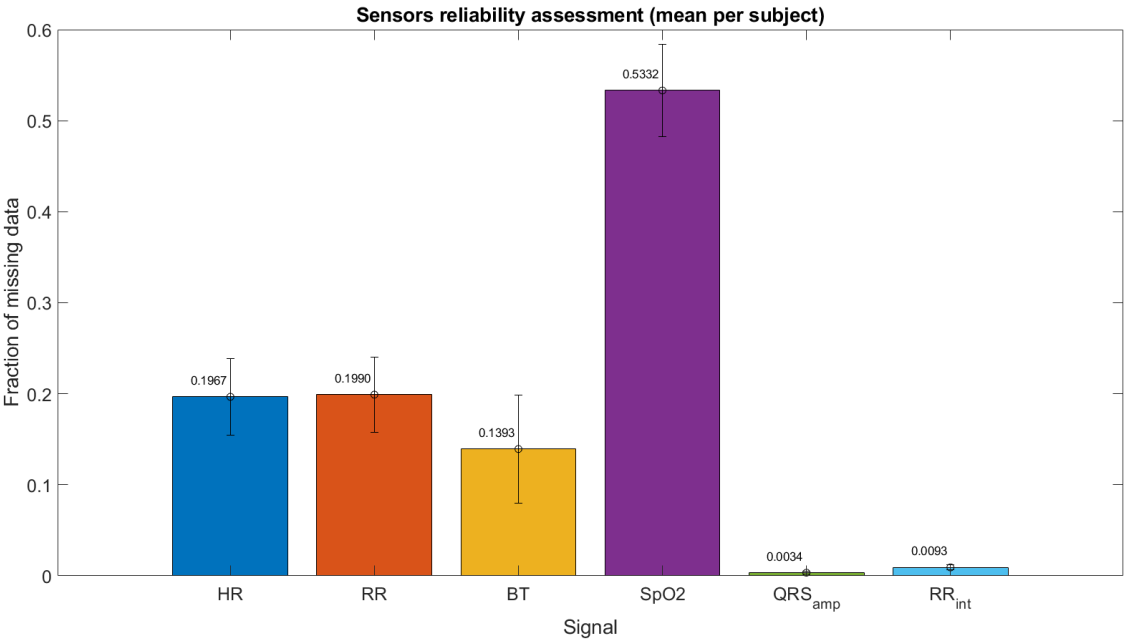
Based on these results it can be concluded that the measure that raises more concerns is SpO_2. It presents 53% of missing data across all subjects data, while the mean percentage of missing data per subject belongs to the interval [48, 58] %, with 95% confidence.

That being said, and by combining this information with the already discussed issue regarding the temperature sensor (see 5.1.1), it was decided to also develop prediction models without relying on any features from BTemp and SpO_2, which constitutes the additional dataset variation.

In summary, 24 datasets variations were tested, which are described in table 6.3.



(a) Sensors reliability assessment across all subjects data. The numbers above the bars represent the fraction of missing data across all subjects data.



(b) Sensors reliability assessment based on the mean fraction of missing data per subject. The numbers above the bars represent the mean fraction of missing data per subject, while the error bars represent the 95% confidence interval.

Figure 6.2: Sensors reliability assessment based on the fraction of data that was missing.

Table 6.3: Summary of datasets variations. LinInt, NApp1 and NApp2 have the same meaning as before (see 5.1.2.2). 'All' refers to the initial feature set described in 6.2, which comprehends 431 dimensions. 'NoTemp&SpO_2' refers to a feature set where all dimensions related to features extracted from the BTemp and SpO_2 time series were excluded. This comprehends 305 dimensions.

Variation	Method to deal with missing data in the vital signs time series	Classes ratio	Initial feature set
1	LinInt	1:50 (original)	'All'
2	LinInt	1:50 (original)	'NoTemp&SpO_2'
3	LinInt	1:10	'All'
4	LinInt	1:10	'NoTemp&SpO_2'
5	LinInt	1:4	'All'
6	LinInt	1:4	'NoTemp&SpO_2'
7	LinInt	1:1	'All'
8	LinInt	1:1	'NoTemp&SpO_2'
9	NApp1	1:50 (original)	'All'
10	NApp1	1:50 (original)	'NoTemp&SpO_2'
11	NApp1	1:10	'All'
12	NApp1	1:10	'NoTemp&SpO_2'
13	NApp1	1:4	'All'
14	NApp1	1:4	'NoTemp&SpO_2'
15	NApp1	1:1	'All'
16	NApp1	1:1	'NoTemp&SpO_2'
17	NApp2	1:50 (original)	'All'
18	NApp2	1:50 (original)	'NoTemp&SpO_2'
19	NApp2	1:10	'All'
20	NApp2	1:10	'NoTemp&SpO_2'
21	NApp2	1:4	'All'
22	NApp2	1:4	'NoTemp&SpO_2'
23	NApp2	1:1	'All'
24	NApp2	1:1	'NoTemp&SpO_2'

6.4 Prediction models development

After having prepared the datasets, prediction models could be trained and developed. Given the nature of the problem in hands (binary classification problem), adequate ML algorithms were explored: Logistic Regression (LR) and Boosted Trees (BT).

The first was chosen since it is one of the simplest ML algorithms, which provides an explainability factor and easier interpretability, which, in its turn, is crucial in a medical context. Additionally, the fact that the predictions results are returned in terms of probabilities is particularly useful for interpreting how sure the model is about each prediction. Also, it is an algorithm that does not present high computational costs.

The second is also a simple and interpretable approach, that had previously been recommended for use in imbalanced datasets problems [79], [144], [145], due to its intrinsic ability to focus on the minority class. In addition to that, ensemble tree-based methods have shown to be the ones that achieve higher performance, when several ML algorithms are tested in the same study [78], [105], [118]. More advantages presented by the BT algorithm were already stated in 2.4.2.2.

6.4.1 Implementation

6.4.1.1 Logistic regression

For the development of LR models, several issues had to be addressed. In particular, the implementation of **feature selection procedures**, the selection of appropriate **performance metrics** to evaluate the models, a **hyperparameter optimization** and the **probability threshold setting** (threshold explained in 2.4.2.1).

Regarding the **feature selection procedures**, three were tested. The first was based on an univariate feature ranking method for classification, that uses chi-square tests of independence between each feature and the output variable to evaluate the feature's importance. The smaller the p-value, the bigger the dependency between feature and output variable and, therefore, more important the feature is. The features were then ordered by importance (in descending order) and a log-likelihood analysis was employed to decide how many features should be kept. This analysis consisted in training a LR model using only the most important feature, calculating the 10-fold cross-validated log-likelihood of the training and validation subsets of the training set and, then, repeating this process iteratively by adding the next most important feature. An example of the results of this analysis applied to the NApp1 dataset for the initial feature set 'All' is shown in figure 6.3. All the other variations presented similar results, so the 30 most important features were always kept, when using this feature selection procedure.

The second procedure relies on a regularization method. Regularization is used when developing LR models to avoid overfitting. Overfitting in LR can be identified by the presence of large regression coefficient values. So, what regularization does is shrinking the coefficient values, which favors less complex solutions with reduced variance, hence decreasing overfitting [79]. This is achieved by adding a term to equation 2.4 that will penalize high regression coefficient values. In lasso regularization, this term is the L_1 norm and the likelihood function to be maximized becomes:

$$pen\_ll(\hat{\beta}) = ll(\hat{\beta}) - \lambda \sum_{j=1}^{n} |\hat{\beta}_j| \tag{6.1}$$

where $\hat{\beta}$ is the current estimation for the regression coefficients, $pen\_ll(\hat{\beta})$ is the penalized likelihood function, $ll(\hat{\beta})$ is the result of equation 2.4, λ is the regularization strength, n is the number of features and $\hat{\beta}_j$ is the current estimation for the regression coefficient associated with feature j.

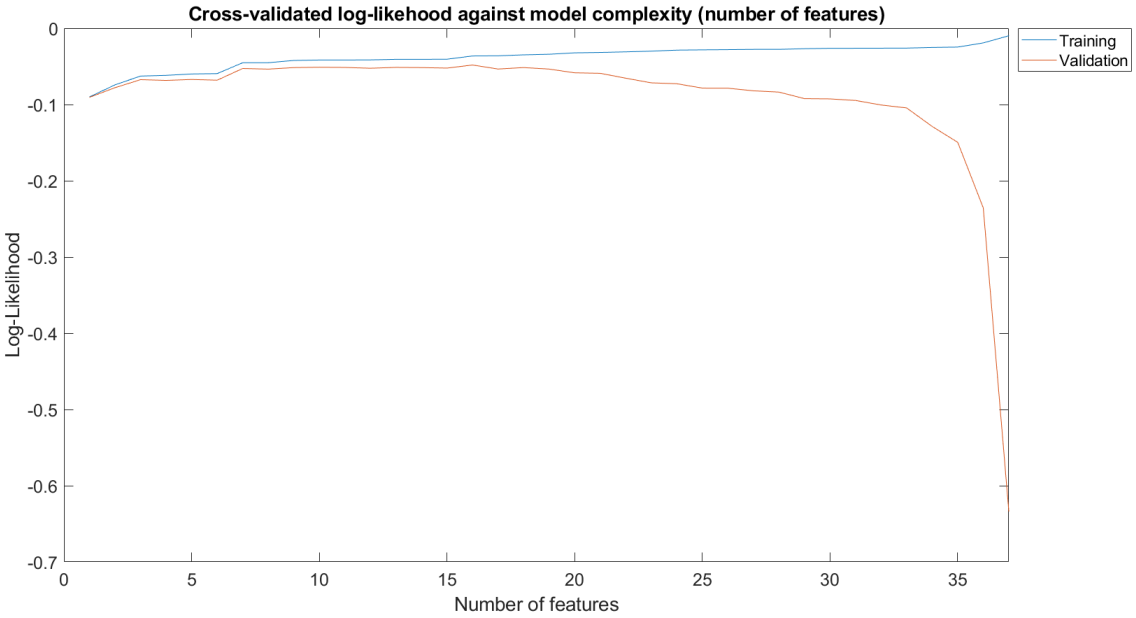


Figure 6.3: Log-likelihood analysis to select an appropriate number of features using the chi-square tests of independence method for feature selection. The region on the left (until around 25 features) represents underfitting situations, since the training and validation sets log-likelihood is similar. The region on the right (after around 35 features) represents overfitting situations, since the training and validation sets log-likelihood is starting to drift apart and the training set log-likelihood is highly tending to 0. That being said, the optimal region is between around 25 and around 35 features, which led to selecting 30 as the suitable number of features. The region after around 35 features is not shown since it represented severe overfitting situations. Note that the mentioned training and validation sets are subsets of the training set, obtained via cross-validation.

Lasso regularization, however, can be used as a feature selection method, not only to reduce overfitting. The reason for this is that the added term in equation 6.1 will cause some regression coefficients to reach exactly zero [79]. This leads to a sparse solution where only the relevant features remain with a non-zero regression coefficient.

To achieve this, a value for the regularization strength, λ, must be chosen. A correct way to estimate the optimal value for this parameter is through cross-validation in the training set [79]. This was performed considering a set of regularization strengths logarithmically spaced and 10-fold cross-validation. Figure 6.4 shows an example of the results of this analysis applied to the NApp1 dataset for the initial feature set 'All'. Instead of maximizing the log-likelihood, this analysis was performed with the goal of minimizing the deviance, which is an equivalent loss function[3]. Lasso tends to favor sparse solutions, hence favoring a smaller λ than the optimal one for feature selection [79]. Consequently, the optimal λ was deemed to be the largest one where the corresponding deviance was within one standard error of the minimal deviance.

The features with associated non-zero regression coefficients, after lasso regularization

[3]in this context, a loss function can be interpreted as the inverse of a likelihood function. This means that instead of maximizing it, the intention is to minimize it

with the optimal λ, were the ones kept, when using this feature selection procedure.

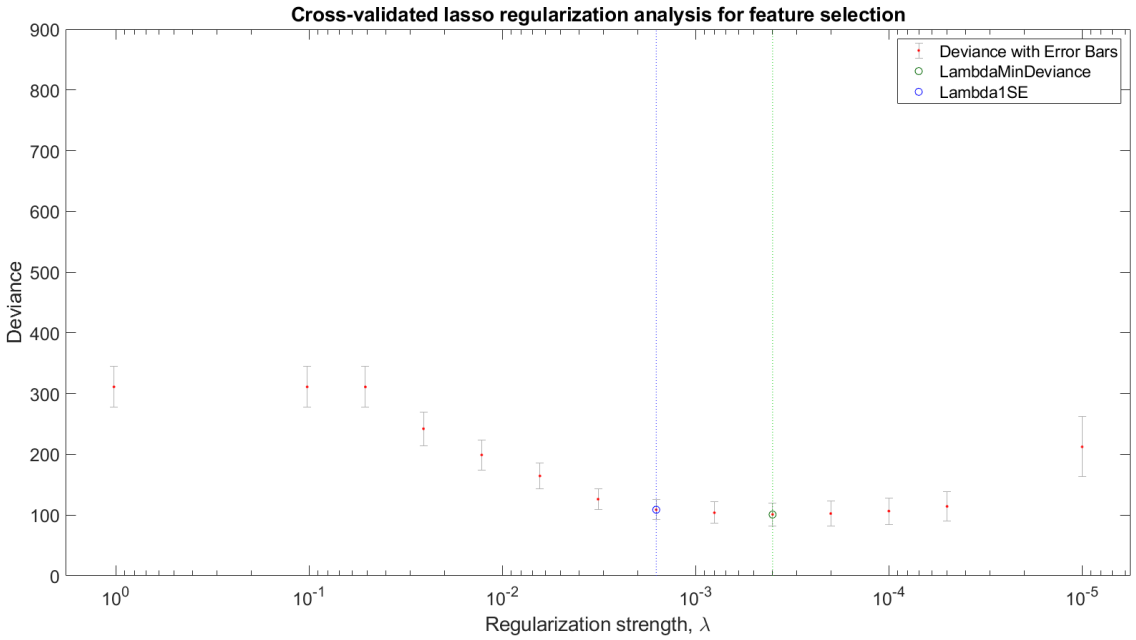


Figure 6.4: Lasso regularization analysis for feature selection. $LambdaMinDeviance$ represents the λ with minimal deviance. $Lambda1SE$ represents the largest λ within one standard error of the minimal deviance, which was considered the optimal one.

The third procedure involved the use of forward and backward stepwise regression. This is an iterative method that adds or removes a feature to a regression model based on the feature's ability to explain the output variable.

In this thesis, this method was started using a constant model (no features included). Then, at each step, a feature would be added or removed to the model constructed in the previous step, based on the Bayesian information criterion, BIC. This is, the change that resulted in a smaller BIC was the one executed at that step. This process would stop when no feature addition or removal would improve the model's performance, as assessed by the BIC, and the features included in the final model were the ones kept.

BIC is a metric that balances the quality of the model fit with the number of features included. Its calculation is given by:

$$BIC = -2 \times ll(\hat{\beta}) + n \times \ln(m) \tag{6.2}$$

where $\hat{\beta}$ is the current estimation for the regression coefficients, $ll(\hat{\beta})$ is the log-likelihood with those coefficients, n is the number of features included in the current model and m is the number of observations in the training set.

Despite this stepwise procedure and the lasso regularization procedure could be performed during the learning process, they were implemented separately. This means that they were implemented only to attain a set of important features and, then, the final models would be trained separately using those sets.

As a final note on the feature selection procedures, it is relevant to mention that the three were implemented for six datasets variations (different initial feature sets and methods to deal with missing data in vital signs) and always considering the original classes ratio.

Regarding the **performance metrics**, goodness of fit and discriminatory metrics were implemented to assess the models performance. These were log-likelihood, AUC, precision, recall, F_1-Score and specificity. All of them were reported considering the test set, so that an unbiased evaluation of the model is provided. Additionally, three performance curves were plotted: EWS efficiency curve, receiver operating characteristic curve and precision-recall curve.

Regarding the **hyperparameter optimization**, this refers to the regularization performed during model development. Unlike the lasso regularization mention before, the regularization discussed now was used to avoid overfitting situations and was included in the model's learning process, not with the goal of selecting features before the learning process.

This type of regularization is designated as ridge regularization. The main difference, when comparing with lasso regularization, is the fact that ridge causes the regression coefficients to tend to zero, but without actually reaching it. This is the case because, for ridge regularization, the term added to equation 2.4 is the L_2 norm. Thereby, the penalized likelihood function to be maximized is given by:

$$pen\_ll(\hat{\beta}) = ll(\hat{\beta}) - \frac{\lambda}{2}\sum_{j=1}^{n}\hat{\beta}_j^2 \tag{6.3}$$

where $\hat{\beta}$ is the current estimation for the regression coefficients, $pen\_ll(\hat{\beta})$ is the penalized likelihood function, $ll(\hat{\beta})$ is the result of equation 2.4, λ is the regularization strength, n is the number of features and $\hat{\beta}_j$ is the current estimation for the regression coefficient associated with feature j.

As before, a value for the regularization strength, λ, must be chosen, which is the hyperparameter being optimized. Once again, this was performed considering a set of regularization strengths logarithmically spaced, using 10-fold cross-validation in the training set and minimizing deviance. However, here, the optimal λ is the one that corresponds to the minimal deviance. Figure 6.5 shows an example of the results of this analysis applied to the NApp1 dataset for the initial feature set 'All' and when considering the final feature set returned by the lasso procedure.

Ridge regularization was used for model development, instead of lasso, because empirically it often results in better predictive performance [79].

Regarding the final issue, the **probability threshold setting** (threshold explained in 2.4.2.1), it was done resorting to the F_1-Score. This is, the threshold was set to be the one that maximized the F_1-Score in the validation set, using 10-fold cross-validation in the training set. The chosen metric to maximize was the F_1-Score, since it represents

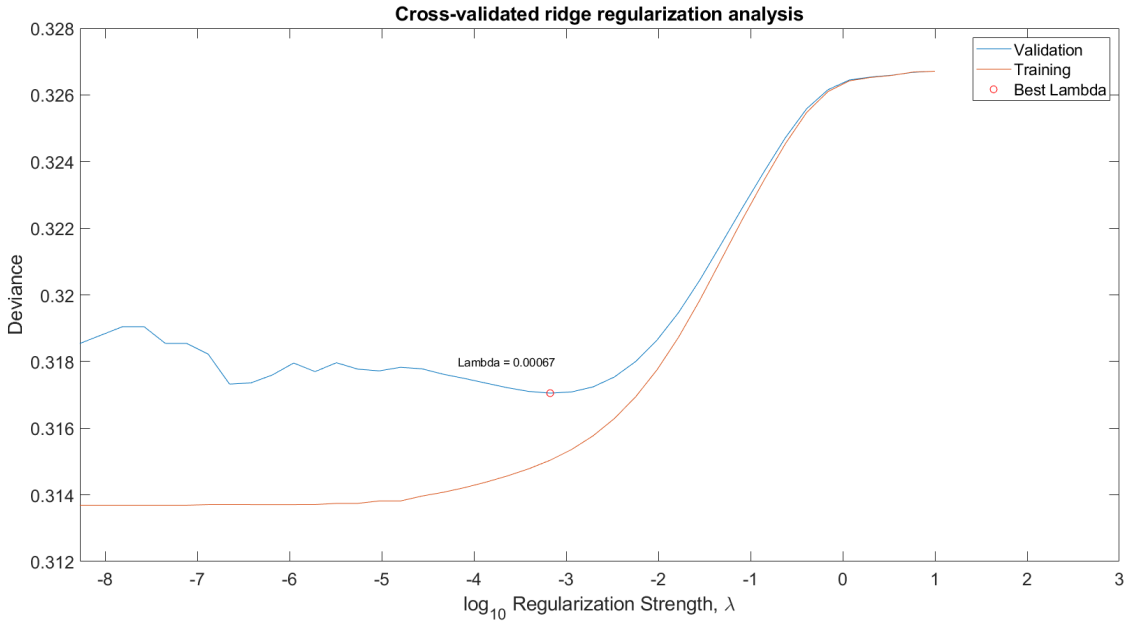


Figure 6.5: Ridge regularization analysis for selection of the optimal regularization strength, λ. The region on the left (until around $\log_{10}(\lambda) = -5$) represents overfitting situations, since the training and validation sets deviance is starting to drift apart and the training set deviance stabilized at a minimum value. The region on the right (after around $\log_{10}(\lambda) = -2$) represents underfitting situations, since the training and validation sets deviance is similar. The optimal λ is identified with a circle and the respective value. Note that the mentioned training and validation sets are subsets of the training set, obtained via cross-validation.

a trade-off between precision and recall, which are the most important metrics in this context, as discussed further ahead in 6.4.2.

By addressing these issues, several LR models could appropriately be developed. These were developed for all the datasets variations and considering the three feature selection procedures, separately. They were all implemented with ridge regularization, using the respective optimal regularization strength.

6.4.1.2 Boosted trees

As for LR models, several issues had to be addressed for the development of BT models. In particular, the implementation of **feature selection procedures**, the selection of appropriate **performance metrics** to evaluate the models, multiple **hyperparameters optimization** and the choice of a **boosting technique**.

Regarding the **feature selection procedures**, in contrast with the LR procedures, the one implemented for BT was intrinsic to the learning process. This implies that the set of important predictors was learned while developing the model.

Since the weak learners in BT are decision trees, at each partition, the locally optimal feature and respective optimal cut-off points, will be selected to partition the feature space. The way these are chosen depends on one of two metrics (depending on the

hyperparameters optimization discussed ahead). The first, the Gini index, is determined by $H_n = 1 - \sum_i p^2(i)$, where $p(i)$ is the fraction of observations with class i in node n, the summation is over all existing classes at node n and H_n is node n's impurity, which can be interpreted as a measure of ability to distinguish between classes in the node (where 0 is maximum ability). The second, entropy, is determined by $H_n = \sum_i -p(i)\log_2 p(i)$, with $p(i)$ and i having the same meaning and H_n having a conceptually similar meaning as before.

Having these metrics defined, when a tree node is to be split, the information gain caused by selecting a particular feature and cut-off points will be calculated as:

$$IG = H_n - \sum_{j=1}^{d} \frac{p_j + n_j}{p + n} H_j \tag{6.4}$$

where IG is the information gain, H_n has the previously described meaning, d is the number of sub-nodes caused by the data partition being tested, p and n are the node n's number of observations from the positive and negative class, respectively, p_j and n_j are the sub-node j's number of observations from the positive and negative class, respectively, and H_j is the sub-node j's impurity.

The feature and cut-off points that result in a higher information gain are considered the optimal data partition at that point. This means that some features might not be even used and only the locally important ones at each split are employed. By summing the information gains a feature provides, over all nodes and all decision trees in the ensemble model, a measure of feature importance is obtained.

Regarding the **performance metrics**, the same as for LR models were implemented, with the exception of the goodness of fit metric. Instead of log-likelihood, the exponential loss was calculated, since the AdaBoost algorithm for binary classification can be interpreted as a minimization of this loss function [83]. The exponential loss is determined by:

$$loss = \sum_{i=1}^{m} \alpha_i e^{-y_i f(\boldsymbol{x_i})} \tag{6.5}$$

where $loss$ is the exponential loss, m is the number of observations in the set being considered, α_i is the weight of observation i, y_i is the true label of observation i (encoded as -1 and 1, instead of 0 and 1) and $f(\boldsymbol{x_i})$ is the predicted score for observation i, which is represented by the feature vector $\boldsymbol{x_i}$.

Regarding the **hyperparameters optimization**, the approach employed for LR could not be utilized here. The reason for this is that, for BT, multiple hyperparameters were to be optimized and individually finding each hyperparameter value that corresponds to the minimum loss function value (as performed for the LR hyperparameter) does not guarantee that the globally minimum loss function value is attained.

Therefore, Bayesian optimization was applied. As any optimization procedure the interest is in finding the minimum of a loss function, within some bounded set of hyperparameters that a model's configuration includes [146]. Since it's computationally

inefficient, or even impossible, to compute the value of the loss function for every possible hyperparameters combination (especially when multiple are being optimized and/or their range of values highly spans), a probability model of the loss function is constructed, when performing Bayesian optimization. This is achieved by first acquiring loss function samples (e.g., by randomly selecting hyperparameters values, training a model and assessing the respective loss function result) and then employing Gaussian processes to model the loss function. Then, this probability model of the loss function is explored using an acquisition function to choose which hyperparameters combination to try next. The acquisition function's goal is to guide the search towards the optimal direction by calculating where a potentially good hyperparameter combination is, based on the current loss function probability model. The acquisition function used in this thesis was the Expected Improvement, where its maximum value corresponds to the hyperparameters combination to try next. Finally, this combination would be used to obtain another loss function evaluation and the probability model would be updated, considering this new sample. This process is completed until a stopping condition is reached (in this case, it was a maximum number of 30 evaluations). The hyperparameters combination that corresponds to the minimum evaluated loss function value is deemed to be the optimal one.

The fact that Bayesian optimization uses information from previous attempts to guide the following search choices presents a tremendous advantage, when comparing with methods such as random search or grid search [146]. A more detailed explanation on Bayesian optimization, Gaussian processes and Expected Improvement can be found in Snoek et al. and Brochu et al. [146], [147].

In this thesis, Bayesian optimization was implemented using 10-fold cross-validation in the training set [79] and to optimize five hyperparameters. Three of which were related to the decision trees construction and the other two concerned the ensemble itself. The former three were the maximum number of splits allowed for each individual tree, the minimum number of observations in a node for an additional split to be made and the split criterion, which could be either the Gini index or entropy. The latter two were the total number of decision trees to be trained and the learning rate, which is a parameter that shrinks the weights of each decision tree, to avoid overfitting.

As a final note on the hyperparameters optimization, it is relevant to mention that this procedure was implemented for six datasets variations (different initial feature sets and methods to deal with missing data in vital signs) and always considering the original classes ratio.

Regarding the choice of a **boosting technique**, the decision fell onto the AdaBoost for binary classification because it preserves the boosting methods advantages and its simplicity provides a higher level of interpretability.

By addressing these issues, several BT models could appropriately be developed. These were developed for all the datasets variations, considering the respective optimal hyperparameters.

6.4.1.3 Comparison reference: MEWS

In order to have a reference model to compare the results obtained with the ML models developed, an EWS was implemented. Given that these are still the standard practice for clinical deterioration detection in general wards, this comparison can highlight how much improvement could be attained by employing more adequate prediction strategies.

As mentioned in 4.2.1, intermittent manual vital signs measurements and MEWS assessments were performed by the nurses, in the MoViSign study. Therefore, MEWS was selected as comparison reference. For MEWS calculation, six measures are required, as demonstrated in table 3.1.

Each set of measures manually acquired was considered an observation. For fair comparison, in the "Event" subjects, observations distanced 12 hours or less from the deterioration event were labeled with 1 (positive class). The remaining and all observations from the "Non-Event" group were labeled with 0 (negative class).

To deal with missing values in the observations, the most recent available value in the previous 12 hours was brought forward, if any exists. This strategy has been used before (see table A.1) and the value of 12 hours was guided by previous research [78], [108], [109].

After handling missing values, if any observation still had less than four measures available (out of six), it would be excluded.

MEWS performance was assessed considering two thresholds: MEWS ≥ 3 and MEWS ≥ 4.

6.4.2 Best models' selection criteria and overall comparisons

Besides the 24 datasets variations, 2 different ML algorithms and 3 different feature selection methods for LR were explored. Therefore, 96 different models were built. Given this elevated number of models, objective and well-defined criteria to select the best ones had to be established.

Thereby, the model's selection procedure laid on the following four criteria:

1. **overfitting/underfitting** - if any model was overfitting or underfitting the training set, it was promptly excluded. These two conditions were assessed by comparing the train and test sets goodness of fit metrics. Additionally, a training AUC equal to 1 was also considered an overfitting situation.

2. **performance metrics** - since the datasets explored in this thesis presented an imbalance between the number of observations of the two classes, accuracy or classification error would be inappropriate and misleading metrics[4] [144], [145]. Therefore, the three chosen metrics to compare models' performance were the test set AUC, precision and recall. The first was used because it's the most widely reported metric

[4] as an example, a simple and naive classifier that would always predict 0 (negative class) would attain around 98% accuracy in the original ratio datasets.

in this context, as demonstrated in table A.1. Precision and recall, on the other hand, were selected for their importance in the presence of imbalanced datasets. Indeed, in this context, these two metrics answer to two crucial questions: recall - "What fraction of the deterioration windows (1-labeled) are correctly predicted by the model?"; precision - "What fraction of the alarms that would be set off (cases where the model predicts positive class) are actually true?". In fact, using these metrics and the F_1-Score as evaluation criteria and the analysis of the precision-recall curve have been suggested for a more adequate representation of model's performance in the presence of imbalanced datasets [145].

3. **model complexity** - if two models were performing similarly, as evaluated by the metrics defined in the previous topic, the least complex model was deemed to be better. This is, the one that required the smaller amount of features would prevail. This criterion was based on the famous principle of Occam's razor.

4. **training set size** - if two models were performing similarly and exhibit identical complexity, the one trained with more data was preferred, since it will probably generalize better for other cohorts. Models with ratio 1:1 were automatically excluded, due to their extremely small training set size.

By applying these criteria, the best model considering the initial feature set 'All' and the best model considering the initial feature set 'NoTemp&SpO_2' were identified. Only the results regarding these two models are discussed, since it was impractical to provide an analysis on the 96 models. These results include a detailed description of the two models and the corresponding performance obtained.

In addition to that, the two above-mentioned models were analyzed in three more aspects:

1. **final set of features** - features importance and the features distribution across the different types of data were reported and discussed.

2. **(mis)classification analysis** - where all windows labeled 1 in the test set, correctly and incorrectly predicted, were analyzed in terms of time before the deterioration event. This allows to investigate how early the model can predict deterioration.

3. **comparison with MEWS** - the discriminatory performance metrics calculated for the ML models were also calculated for the comparison reference, MEWS, and compared with those models results. Also, the three performance curves, mentioned in this thesis, were also plotted for MEWS, for further comparison.

In addition to selecting the best models, **overall comparisons** were performed. This is, given that so many models variations were tested, they could be grouped by the different characteristics, and differences in performance between each characteristic variations could be assessed. The LR feature selection method characteristic was not considered for

these analysis. This means that the feature selection method that yielded better results for a given variation was the one considered and that 48 models variations were included for the overall comparisons (24 datasets variations × 2 ML algorithms).

As an example, if we group the models by the initial feature set, table 6.4 will be obtained. Each of those models will have an associated AUC and F_1-Score. Therefore, hypothesis tests can be applied to assess if the median or mean AUC and F_1-Score, for the initial feature sets 'All' and 'NoTemp&SpO_2', are statistically different. This was performed using the paired Wilcoxon signed rank test at 5% significance level, and grouping by the four different characteristics (method to deal with missing data in the vital signs time series, classes ratio, initial feature set and model type). The results of this analysis are discussed in 6.4.3.2, with the exception of the classes ratio grouping. The reason for this is that no actual conclusions could have been drawn, since, when undersampling, two other important factors might play a role. These are the smaller dataset size and the fact that data complexity can change [144].

The goal of these overall comparisons was to determine which was the best variation for each of the characteristics. For example, if NApp2 would show statistically significant better performance metrics than LinInt, it probably is a superior technique to handle missing data in vital signs, at least for the deterioration prediction task in this context.

6.4.3 Results and discussion

6.4.3.1 Best models

The best model, when considering the initial feature set 'All', was a LR model regularized with $\lambda = 5.5 \times 10^{-6}$, which corresponded to the optimal regularization strength. The feature selection method employed in this model's construction was the stepwise regression procedure, which resulted in only 9 features being deemed relevant. The dataset used for developing this model was from the LinInt type and presented a classes ratio of 1:10.

Regarding the best model with initial features set 'NoTemp&SpO_2', this was a LR model regularized with $\lambda = 1.4 \times 10^{-4}$, which corresponded to the optimal regularization strength. The feature selection method employed in this model's construction was the lasso regularization procedure, which resulted in 42 features being deemed relevant. The dataset used for developing this model was from the NApp2 type and presented a classes ratio of 1:10.

Performance metrics

The performance achieved by the two models is reported in table 6.5. Additionally, figures 6.6, 6.7 and 6.8 provide more insight on the models' performance, by displaying the models' receiver operating characteristic curves, EWS efficiency curves and precision-recall curves, respectively. The figures also include the same curves for the comparison reference, MEWS.

Table 6.4: Example of how the different models variations would be grouped for the overall comparisons. The characteristic being considered for the grouping is the initial feature set. LinInt, NApp1 and NApp2 have the same meaning as before (see 5.1.2.2). 'All' refers to the initial feature set described in 6.2. 'NoTemp&SpO_2' refers to a feature set where all dimensions related to features extracted from the BTemp and SpO_2 time series were excluded.

Initial feature set	Method to deal with missing data in the vital signs time series	Classes ratio	Model type
'All'	LinInt	1:50 (original)	LR
'All'	LinInt	1:50 (original)	BT
'All'	LinInt	1:10	LR
'All'	LinInt	1:10	BT
'All'	LinInt	1:4	LR
'All'	LinInt	1:4	BT
'All'	LinInt	1:1	LR
'All'	LinInt	1:1	BT
'All'	NApp1	1:50 (original)	LR
'All'	NApp1	1:50 (original)	BT
'All'	NApp1	1:10	LR
'All'	NApp1	1:10	BT
'All'	NApp1	1:4	LR
'All'	NApp1	1:4	BT
'All'	NApp1	1:1	LR
'All'	NApp1	1:1	BT
'All'	NApp2	1:50 (original)	LR
'All'	NApp2	1:50 (original)	BT
'All'	NApp2	1:10	LR
'All'	NApp2	1:10	BT
'All'	NApp2	1:4	LR
'All'	NApp2	1:4	BT
'All'	NApp2	1:1	LR
'All'	NApp2	1:1	BT
'NoTemp&SpO_2'	LinInt	1:50 (original)	LR
'NoTemp&SpO_2'	LinInt	1:50 (original)	BT
'NoTemp&SpO_2'	LinInt	1:10	LR
'NoTemp&SpO_2'	LinInt	1:10	BT
'NoTemp&SpO_2'	LinInt	1:4	LR
'NoTemp&SpO_2'	LinInt	1:4	BT
'NoTemp&SpO_2'	LinInt	1:1	LR
'NoTemp&SpO_2'	LinInt	1:1	BT
'NoTemp&SpO_2'	NApp1	1:50 (original)	LR
'NoTemp&SpO_2'	NApp1	1:50 (original)	BT
'NoTemp&SpO_2'	NApp1	1:10	LR
'NoTemp&SpO_2'	NApp1	1:10	BT
'NoTemp&SpO_2'	NApp1	1:4	LR
'NoTemp&SpO_2'	NApp1	1:4	BT
'NoTemp&SpO_2'	NApp1	1:1	LR
'NoTemp&SpO_2'	NApp1	1:1	BT
'NoTemp&SpO_2'	NApp2	1:50 (original)	LR
'NoTemp&SpO_2'	NApp2	1:50 (original)	BT
'NoTemp&SpO_2'	NApp2	1:10	LR
'NoTemp&SpO_2'	NApp2	1:10	BT
'NoTemp&SpO_2'	NApp2	1:4	LR
'NoTemp&SpO_2'	NApp2	1:4	BT
'NoTemp&SpO_2'	NApp2	1:1	LR
'NoTemp&SpO_2'	NApp2	1:1	BT

Table 6.5: Performance metrics obtained for the best model with initial features set 'All' and 'NoTemp&SpO_2'. The discriminatory metrics are all reported considering the test set.

Model	Log-likelihood (Train/Test)	AUC [95% confidence interval]	Recall	Precision	F_1-Score	Specificity
'All'	-0.068/-0.116	0.97 [0.81, 1]	0.77	0.83	0.80	0.99
'NoTemp&SpO_2'	-0.121/-0.130	0.94 [0.79, 1]	0.85	0.79	0.82	0.98

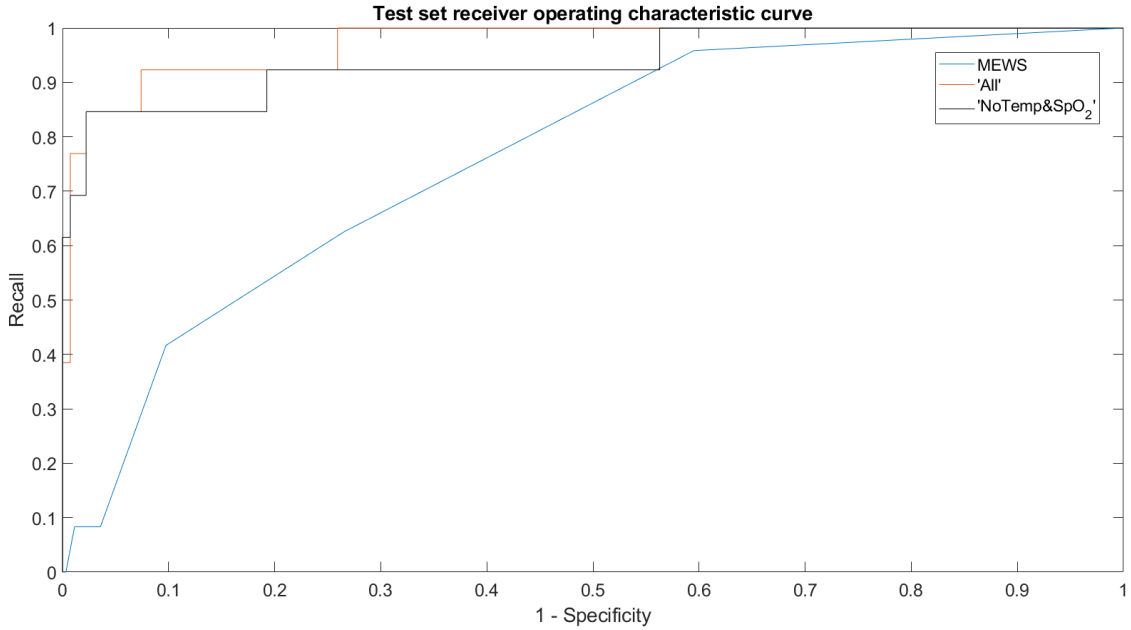


Figure 6.6: Receiver operating characteristic curves for the best model with initial feature set 'All' and 'NoTemp&SpO_2'. Additionally, the respective curve for the comparison reference, MEWS, is displayed ($MEWS$).

The precision attained by the two models corresponds to stating that 83% and 79%, respectively, of the alarms that would be set off by the model, would actually be true. This is an incredible improvement when compared with the result obtained by MEWS (see table 6.7). In fact, many ML-based models have already been used to aid in decreasing the rate of false alarms in clinical settings, despite this still being an unresolved problem [5]. Proof of that is the fact that, in the reviewed work, only Tarassenko et al. [101] achieved a satisfactory result, reporting a 95% precision with their novelty detection algorithm. On the other hand, Mao et al. [108] and Zimlichman et al. [114] only managed to reach 30% and 54% precision, respectively. The remainder of the studies did not report this metric, despite its importance.

These studies' results and a precision of 83% and 79% are, then, indicative that this more advanced and complex algorithms might be the way to address one of the existing limitations of EWS, the elevated rate of false alarms. Indeed, such improvement, as the

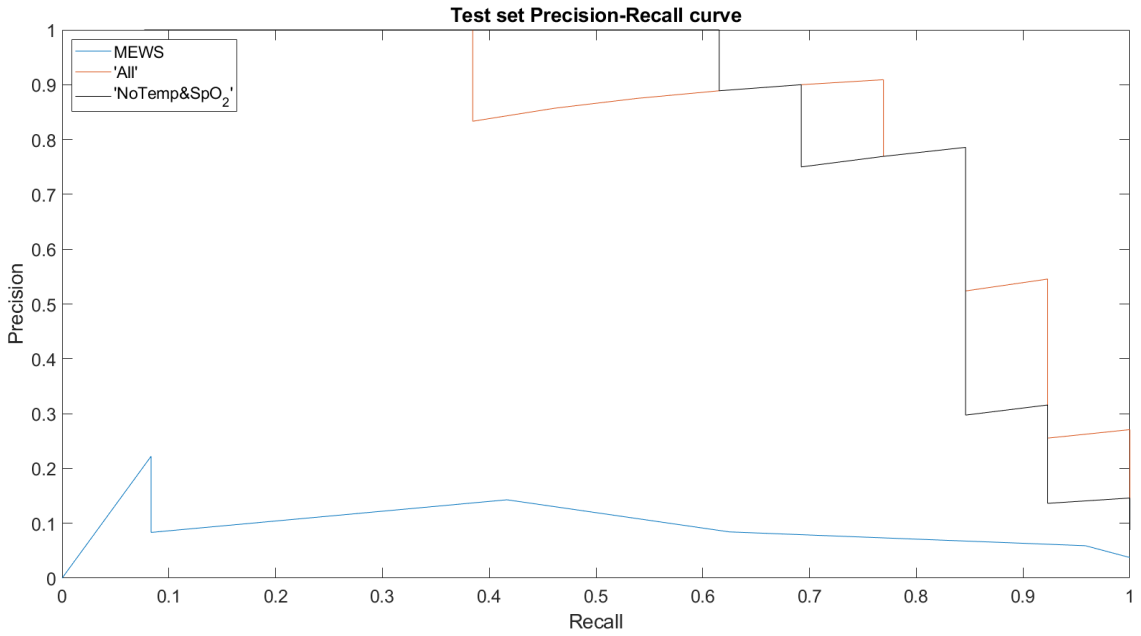


Figure 6.7: Precision-recall curves for the best model with initial feature set 'All' and 'NoTemp&SpO_2'. Additionally, the respective curve for the comparison reference, MEWS, is displayed ($MEWS$).

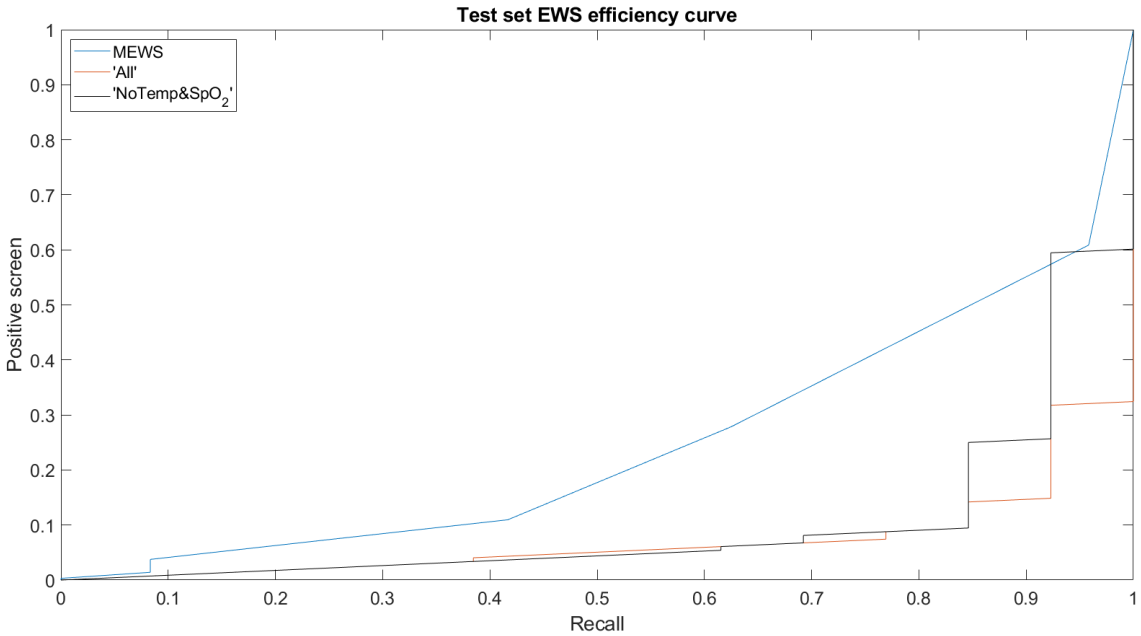


Figure 6.8: EWS efficiency curves for the best model with initial feature set 'All' and 'NoTemp&SpO_2'. Additionally, the respective curve for the comparison reference, MEWS, is displayed ($MEWS$). Positive screen represents the proportion of observations classified as positive by the model.

one obtained here, might significantly help reduce alarm fatigue in clinical staff. However, these results still imply that, approximately, one in five alarms are false, which may yet reveal excessive.

Regarding the recall accomplished by the models, it corresponds to saying that 77% and 85%, respectively, of the deterioration windows analyzed were correctly predicted by the model. Once again, this represents a remarkable enhancement facing MEWS results (see table 6.7). In medical contexts, high recall is strongly required [10], as the not identification of a deteriorating patient might have serious consequences for its health status. Indeed, in this context, most studies focus on achieving high recall and/or specificity. In the reviewed literature, were found recalls between 41% and 100%, whereas only Clifton et al. [97] and Zimlichman et al. [114] algorithms relied on continuous monitoring. The former reported 96% recall, when predicting deterioration with one hour in advance. The latter obtained recalls between 55% and 100%, but their system requires the patient to be in bed, not allowing the monitoring of ambulatory patients.

Despite the recall results obtained here being comparable with previous studies and superior to MEWS, they still mean that, approximately, one in each four or six assessments performed by the warning system, that should be regarded as deterioration-related, is deemed as coming from a currently healthy subject. The repercussions of this situation are various, and were already mentioned before (see 1.1).

With respect to specificity, 99% and 98% values were achieved. Thus, nearly every window that was not followed by deterioration events was correctly predicted. This result is slightly better than the one obtained by MEWS (see table 6.7). Such elevated values, for both the developed models and MEWS, can be justified by the severe class imbalance and small dataset size [148]. Also, in such situations, it is expected that the model will be particularly proficient in predicting the negative class, since it was mostly trained with examples of this class. Thereby, the number of true negatives is expected to be very high, when comparing with the false positives, highly skewing the ratio between them and, hence, increasing specificity.

As stated before, in this context, many studies strive to achieve high specificity. This is related with the fact that this metric depicts the model's ability to not mistake a patient that is healthy as being deteriorating. Specificities between 64% and 95% were reported in the reviewed literature. Yet again, only Clifton et al. [97] and Zimlichman et al. [114] developed models based on continuous monitoring, where specificities of 93% and between 64% and 94% were obtained, respectively. Additionally, Mao et al. [108] stated that for practical implementation, at least 95% specificity is required in hospitals. However, only their model and the ones developed in this thesis fulfilled that condition. Although, in their case, that was at the cost of low recall and precision.

In reality, it is difficult to combine high recall, precision and specificity in the same model. As typically high recall is essential, precision and/or specificity end up being neglected [114]. This fact highlights the results obtained here, since high values for the three metrics were achieved for both models. Not only that, when comparing with MEWS,

it is evident that considerable improvements can be made to the currently employed systems for the assistance of nurses, in the task of detecting clinical deterioration events in surgical patients.

Commonly, those EWS-based systems performance is also evaluated examining the AUC. In fact, the majority of studies that develop new strategies, which intend to replace those systems, focus the performance comparison on this metric, as can be ascertained by inspecting table A.1.

In the same table, AUC values between 0.60 and 0.94 can be found. However, from those who reported it, only Zimlichman et al. [114] model depended on continuous monitoring. They obtained AUC between 0.69 and 0.93, which are still below the 0.97 and 0.94 values achieved by the LR models developed in this study. The underlying reason for such high AUC can be explained by the fact that receiver operating characteristic curves provide an excessively optimistic perspective on an algorithm's performance in the presence of imbalanced datasets [149]. As a matter of fact, the interpretability of this curve and, hence, AUC, have been reported to be misleading in these situations, with respect to possible conclusions about the model's performance [148]. Therefore, despite the exciting results obtained for AUC and specificity, these should be interpreted with caution and the focus should mainly be on precision and recall, as advised in prior research [145].

Also, in the presence of imbalanced datasets, it is suggested the report and analysis of the precision-recall curve, instead of the receiver operating characteristic curve, since it can more properly represent the model's performance [145], [148]. For this reason, no further discussion will be performed on the receiver operating characteristic curves.

On the other hand, the precision-recall curves, shown in figure 6.7, are now discussed.

A model's performance, when judged by the precision-recall curve, is as positive as its curve proximity to the top right corner of the precision-recall space [145]. Hence, it can be concluded that the LR models present a remarkably superior performance than MEWS. Indeed, MEWS skill is clearly unacceptable across all threshold values, for deployment in hospital context, since its precision is never higher than 22%. This restates how much improvement can be attained in the prediction task, by employing more adequate and personalized models instead of simple and general EWS.

The last curves plotted, figure 6.8, display the obtained EWS efficiency curves. These curves depict the rate of positive predictions (positive screen) made by the models, in order to attain a certain value of recall. If we fix the discussion on the recall of the best model with initial features set 'All', 77%, one can conclude that MEWS would have to set off 33% (42% vs 9% positive screen) more alarms than the developed model. To put this number in perspective, considering the total number of available windows in the original dataset (2152), this would represent around 710 more alarms being triggered. Yet, if we consider that, with continuous monitoring, patient's assessments can be performed hourly (see discussion in 6.5.2), this result implies that, approximately, fewer 8 alarms per day per patient would be activated. This has strong implications on the reduction of

alarm fatigue in clinical staff, which is a crucial issue that affects healthcare quality and is yet to be resolved [5]. The obtained disparities in the number of triggered alarms are even more evident for the best model with initial features set 'NoTemp&SpO_2'. At this model's recall, MEWS would have to set off 41% (50% vs 9% positive screen) more alarms. It is also visible that both LR models present the same positive screen for the respective recall (9%).

Finally, the two developed models can be compared with each other. Regarding performance, they can be considered identical, as evidenced by (1) the overlap in the AUC mean confidence intervals; (2) the similar trade-off between precision and recall, i.e., F_1-Score, and similar specificity; (3) the resemblance between the performance curves.

Furthermore, both models were trained with the same amount of data (1:10 vs 1:10 classes ratio). However, the best model with initial feature set 'All' requires only 9 features, while the best model with initial feature set 'NoTemp&SpO_2' requires 42 features. This shows that much more predictors are needed to achieve the same performance in the absence of BTemp and SpO_2 data. Thereby, BTemp and SpO_2 sensors provide relevant, but not necessarily essential, information for the prediction of deterioration events, since similar performance can be attained without them. This brings an additional practical advantage, since a good-performance model can still be employed while patients are required to wear only one sensor.

However the two best models were analyzed, only one could be used for the final warning system assembly. The discussion performed above revealed that the two models perform similarly and were trained with the same amount of data. However, the model with the initial feature set 'All' requires a smaller number of features, which by the third criteria for model selection would make it the preferred choice. Nonetheless, due to the formerly discussed BTemp and SpO_2 sensors unreliability, it was decided to assemble the final warning system using the best model considering the initial feature set 'NoTemp&SpO_2'.

Features analysis

The features analysis regarding the best model with initial features set 'All' is very superficial, for brevity reasons and given that this was not the model used in the final warning system.

As mentioned before, that model utilized 9 features. No correlations between them were present, which avoids multicollinearity issues. Two of the features were personalized ($RR\_normdiff$ and $HR\_catcoef$) and one explored correlations between vital signs (correlation between HR and RR). Table 6.6 presents the features distribution across the different types of data explored. As expected [18], [54], [78], [103], [104], [117], HR and RR were the signals that contributed to more features.

A deeper and more complete discussion is now performed on the final set of features employed by the model used in the final warning system, the best model with initial

Table 6.6: Features distribution across the different types of data explored, for the best model with initial feature set 'All' and 'NoTemp&SpO_2'. This is reported as absolute number of features extracted from the respective type of data, and as a percentage of the total number of features.

Model	HR	RR	BTemp	SpO_2	QRSa	RRI	Demographic	**Total**
'All'	3 (33.(3)%)	3 (33.(3)%)	2 (22.(2)%)	0 (0%)	2 (22.(2)%)	0 (0%)	2 (22.(2)%)	**9**[*]
'NoTemp&SpO_2'	13 (30.1%)	16 (38.1%)	—	—	5 (11.9%)	6 (14.3%)	8 (19.0%)	**42**[**]

[*] one feature involved both HR and RR. Two features involved both one of the continuous signals and demographic information.
[**] one feature involved both HR and RR. Five features involved both one of the continuous signals and demographic information.

features set 'NoTemp&SpO_2'. Features importance was assessed based on the regression coefficients absolute values. The reason for this is that, in this context, the regression coefficients provide a measure of how a change in a feature's value relates to changes in the probability of predicting an assessment as deterioration-related.

In figure 6.9, the 15 most important dimensions are displayed. Additionally, table C.1 lists, by descending importance, the 64 dimensions included for this model development. These are the result of the extraction of the 42 features deemed relevant by the lasso regularization procedure. This table highlights that not all odds ratio[5] make physiological sense, which might be a consequence of the small dataset size. Thereby, if the exact same methodology employed for this model development, would be employed with a larger and more representative dataset, it would be expected that, at least, some of the regression coefficients re-estimations would widely differ from the current configuration.

Regarding the 42 final features, no correlations between them were present, which avoids multicollinearity issues. However, the lasso feature selection procedure tends to choose arbitrarily between correlated features [79]. Therefore, table C.1 indicates which non-utilized features were correlated with the ones that were used in the model, since these could have easily been selected instead of the adopted one.

Additionally, the set of 42 features included eight personalized features ($HR\_catcoef$, $RR\_catcoef$, $pd7\_coef$, $pd9\_coef$, num_comorb, $HasMultipleComorbs$, age_coef and $RR\_normdiff$) and one that explored correlations between vital signs (correlation between HR and RR). Plus, from the 15 most important dimensions, nine either include some degree of personalization or involved the exploration of correlations between vital signs.

Both the lack of personalization and exploration of correlations between vital signs have been appointed as limitations presented by the currently employed clinical deterioration detection methods, and by many of the new strategies reviewed in chapter 3. And the results obtained for this model and for the best model with initial features set 'All'

[5] odds ratio represents the odds of predicting positive class in the presence of one unit increase in the respective feature, compared with the odds of predicting positive class in the absence of one unit increase in the respective feature [77]

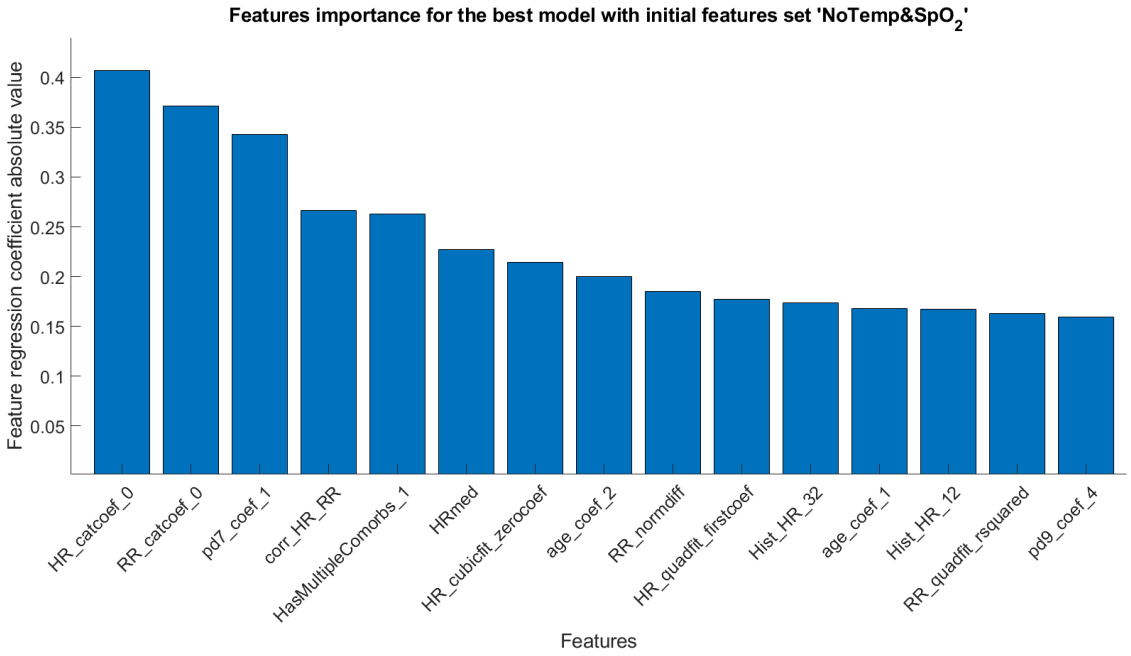


Figure 6.9: Features importance in the best model with initial features set 'NoTemp&SpO_2', based on the associated regression coefficients absolute values. Only the 15 most important dimensions are displayed. These are calculated accordingly to B.2.2, B.2.2, B.2.3, B.1.13, B.2.6, B.1.7, B.1.9, B.2.1, B.1.2, B.1.9, B.1.15, B.2.1, B.1.15, B.1.9 and B.2.3, respectively.

evidence why those are limitations. The fact that so many of the features that comprise these models were the result of exploring these two domains, demonstrates their added value to the deterioration prediction task. Particularly, the inclusion of demographic and contextual information in the prediction model have been advised before [5] and suggested to enhanced performance [11], [98]. Also, this is not the first time that demographic and contextual features were reported to be important predictors for these warning systems [78], [104]. Hence, the integration of personalized features and the analysis of correlations between vital signs is highly suggested.

The notable presence of demographic and contextual features in the final models is emphasized in table 6.6, since it demonstrates that these features contribute for 19% and 22% of the total number of features used in each model. Also, this table evidences that RR is the signal that contributes to more features, followed by HR.

These results are consistent with findings from previous studies in this area [18], [54], [78], [103], [104], [117]. For example, Fieselmann et al. [117] reported that RR was the most important predictor of cardiopulmonary arrest in hospital wards, while Churpek et al. [78] found RR and HR to be the most important predictors in their random forest model. Plus, Kamio et al. [18] reaffirmed that changes in RR are the earliest sign of patient deterioration and that this measure can even predict fatal deterioration events. Finally, Cretikos et al. [54] reported numerous papers that stressed the significance of adequately monitoring RR for the early detection of deterioration events.

A final note can be made to draw the attention to the fact that QRSa contributed to 11.9% and 22.2% of the features in the best models. Considering that this signal had never been used before in this context, a novel important predictor of deterioration might have been found here. Although, given the small cohort studied, this statement requires further validation.

(Mis)classification analysis

Besides discussing the percentage of deterioration windows (1-labeled) that were correctly predicted by the models, i.e., their recall, it is relevant to discuss how long before the deterioration event onset were those windows.

On average, the models correctly predicted deterioration events with 7 ± 3 hours in advance[6] (best model with initial features set 'All') and 7 ± 3 hours in advance[7] (best model with initial features set 'NoTemp&SpO_2'). From the reviewed studies that employed the same prediction strategy (discrete time analysis) [78], [103], [104], [108], [109], none reported this result. Therefore, no direct comparison can be performed. Nonetheless, their strategies involved attempting to predict deterioration events between 4 and 30 hours in advance.

Due to the employed prediction strategy, the developed models' goal was to predict deterioration with 12 hours or less in advance. So, these would never be able to predict earlier than some of the reviewed methods, simply due to their inherent prediction strategy. Also, the result achieved depends on the available deterioration windows distance from the respective deterioration event. In fact, in a scenario where all of those windows would be correctly predicted by the models, the results would have been 8 ± 3 hours[8] of average prediction in advance.

Even so, it could be stated that these models cannot predict deterioration earlier than most of the reviewed models. Although, the achieved result is rather encouraging, since the models developed here are the only fully independent of nurse's manual measurements. Thus, they provide the opportunity for a more frequent patient assessment, which enhances the probability of identifying deterioration in its early stages, without even interfering with nurse's workflow.

Also, it is hard to say whether 7 ± 3 hours is enough for a timely intervention, given the wide variety of deterioration events that can occur in surgical patients. This diversity is evidenced by the amount of distinct types of deterioration events that occurred in such a small study as this one (see table 4.2). Nevertheless, the combination of such an early detection with the much higher performance metrics and with the chance of monitoring patients on a much more regular basis, without requiring manual measurements to be

[6] median 7.50 (6.25/8.75) hours (first quartile / third quartile)
[7] median 8.0 (6.5/9.5) hours (first quartile / third quartile)
[8] median 8 (7/10) hours (first quartile / third quartile)

performed, suggests that these models are much more practical and appropriate than EWS, which can enhance patient's outcomes.

Comparison with MEWS

The obtained performance metrics for MEWS are shown in table 6.7. In addition to that, its performance curves are plotted in figures 6.6, 6.7 and 6.8, alongside with the respective curves for the best model with initial features set 'All' and 'NoTemp&SpO_2'.

Table 6.7: Performance metrics obtained for MEWS. $MEWS3$ and $MEWS4$ refer to MEWS calculation considering the threshold as 3 and 4, respectively.

EWS	AUC [95% confidence interval]	Recall	Precision	F_1-Score	Specificity
MEWS3	0.76 [0.70, 0.82]	0.42	0.14	0.21	0.90
MEWS4	0.76 [0.70, 0.82]	0.08	0.08	0.08	0.96

By doing a similar analysis, as for the best models with initial features set 'All' and 'NoTemp&SpO_2', it is concluded that, with the best MEWS threshold configuration, only 14% of the triggered alarms would actually be true and only 42% of the deterioration-related assessments would be correctly predicted. With the other threshold configuration, these results decay even more, both to 8%.

This illustrates the inability of EWS to predict events so physiologically complex as the deterioration events experienced by these patients. This is not the first time that this EWS ineptitude is reported. In fact, Gao et al. [93] reviewed many of these scores and also found little evidence of reliability, poor recalls and poor predictive value, which is consistent with the findings made here. Therefore, one may ask why these are still the current practice for deterioration detection in general wards. Especially when several new strategies have proved to attain much superior performance than EWS [78], [103], [104], [116]. For example, Churpek et al. [78] tested numerous ML algorithms and all of them attained better performance than MEWS. Escobar et al. [104] models not only achieved higher AUC but also reported a lower rate of false alarms. These results are all in conformity with this study's results, as evidenced by comparing tables 6.5 and 6.7. So, it becomes evident that considerable performance improvements can be attained by employing more adequate prediction models. This change can contribute to better patient's outcomes, by providing nurses with a proper assistance tool, in the task of detecting clinical deterioration events.

The comparison of tables 6.5 and 6.7, and, in particular, the comparison of AUC and specificity can also help justify why, in the presence of imbalanced datasets, these two metrics are not the most appropriate ones to evaluate the models. If the discussion would be focused on them, it could be stated that MEWS performance is, at least, acceptable,

when compared with the LR models. However, as seen before, for deployment in hospital settings, MEWS results are evidently unsatisfactory when assessed by precision and recall.

One final note on MEWS performance can be made to point out that the respective AUC values obtained here are in line with what was reported by other studies [78], [103], [104], [116].

6.4.3.2 Overall comparisons

Vital signs missing data approach

Since there were three different approaches being tested to deal with missing data in the vital signs time series, 16 models had been built considering each of them (48 models variations ÷ 3 approaches). Their performance results (AUC and F_1-Score) were grouped and analyzed considering that partition, and table 6.8 was constructed.

To assess if the differences in performance, between the different approaches, were statistically significant, hypothesis tests were performed and the results obtained are displayed in table 6.9.

Table 6.8: Summary of the models performance metrics, when grouping the results by the approach to deal with missing data in the vital signs time series.

Approach	AUC (mean ± SD)	AUC (median [first quartile / third quartile])	F_1-Score (mean ± SD)	F_1-Score (median [first quartile / third quartile])
NApp1	0.85 ± 0.07	0.84 [0.81 / 0.91]	0.7 ± 0.1	0.68 [0.62 / 0.78]
NApp2	0.91 ± 0.04	0.92 [0.88 / 0.94]	0.79 ± 0.08	0.80 [0.75 / 0.82]
LinInt	0.89 ± 0.06	0.90 [0.85 / 0.93]	0.7 ± 0.1	0.77 [0.61 / 0.80]

Table 6.9: Results of the hypothesis tests performed to assess if the differences in the performance metrics were statistically significant, between the different approaches to deal with missing data in the vital signs time series.

	AUC			F_1-Score		
	NApp1 vs NApp2	NApp1 vs LinInt	NApp2 vs LinInt	NApp1 vs NApp2	NApp1 vs LinInt	NApp2 vs LinInt
p-value	< 0.01	0.06*	< 0.05	< 0.01	0.17*	< 0.01

* not significant

The combination of results from table 6.8 and 6.9 indicates that the new approach version 2 consistently performed better, both in terms of AUC and F_1-Score, than the new approach version 1 and linear interpolation. No significant difference in performance was found between the new approach version 1 and linear interpolation.

These results reinforce the idea that the inclusion of personalized methods in these warning systems can indeed enhance performance in the early detection of deterioration

events, as hypothesized before [5], [11]. This was also confirmed by Clifton et al. [98], that have shown that deterioration was more accurately predicted earlier, when vital signs time series were corrected using their personalized framework based on Gaussian processes. Both Clifton's [98] results and the ones obtained here demonstrate that handling periods of missing data, in the vital signs time series, with an appropriate and personalized method, instead of the commonly used simple and generic approaches, is a strategy that most likely will have a beneficial impact in the deterioration prediction task, hence contributing for better patients outcomes.

Nonetheless, linear interpolation also presented some good properties. Despite not significant, it attained better performance than new approach version 1 and a performance close to the one obtained by new approach version 2, despite the significant differences. Furthermore, it was the strategy with smaller relative error rate, across all gap durations tested, and it is computationally faster than the two versions of the new approach (see figure C.19).

Therefore, in spite personalized approaches are recommended, both approaches are valid, and the best one might depend on the study's characteristics and conditions. This is evidenced by the fact that the best model with initial features set 'All' was from the LinInt type, while the best model with initial features set 'NoTemp&SpO_2' was from the NApp2 type.

Model type

Since there were two different ML model types being tested, 24 models had been built considering each of them (48 models variations ÷ 2 model types). Their performance results (AUC and F_1-Score) were grouped and analyzed considering that partition, and table 6.10 was constructed. Additionally, this table also presents the results of the hypothesis tests performed to assess if the differences in performance, between the two model types, were statistically significant.

Table 6.10: Summary of the models performance metrics, when grouping the results by the model type. Both the differences in AUC and F_1-Score, between the two model types, are statistically significant, as demonstrated by the respective p-value.

Model type	AUC (mean ± SD)	AUC (median [first quartile / third quartile])	F_1-Score (mean ± SD)	F_1-Score (median [first quartile / third quartile])
LR	0.91 ± 0.05	0.93 [0.88 / 0.95]	0.78 ± 0.07	0.80 [0.75 / 0.82]
BT	0.85 ± 0.06	0.86 [0.81 / 0.90]	0.7 ± 0.1	0.67 [0.59 / 0.77]
p-value		< 0.01		< 0.01

Indeed, significant differences were found, which indicates that LR consistently performed better than BT, both in terms of AUC and F_1-Score. Also, the fact that the two

best models, previously analyzed, were of the LR type was already a sign that this might have been the case.

These results, however, are contradictory with prior research, which suggested the performance of tree-based models to be superior to that of regression models, based on tests performed across numerous datasets from different fields of study [150]. Besides that, in this context, Churpek et al. [78] found random forests and BT to be the most accurate ML models for clinical deterioration detection in wards, outperforming LR, while Pirracchio et al. [105] achieved the same conclusion when attempting to predict mortality in ICU patients. Chen et al. [118] also reported that random forests outperformed LR models, when attempting to discern between real and artifact vital signs alerts. Nonetheless, some studies [108], [151] reported the opposite, in concordance with the present study's outcome. These discordant findings reflect the fact that no algorithm can be considered the absolute best across all possible scenarios and datasets [78].

Particularizing for this study, the justification for BT worse performance is related to the small dataset size. Previous research [85] have shown that these modern ML algorithms require as much as 10 times more data than classical techniques, such as LR, to produce stable results. In fact, their use in medical prediction problems have been suggested to be performed, only if very large datasets are available [85].

Hence, the low amount of existent data for this thesis development, did not allow the true predictive power of BT to be explored and stable BT models to be produced. However, even in the presence of such instability, BT managed to achieve good test set results, despite probably not generalizable. Thereby, in a future larger study with similar goals, this should be a ML algorithm to take into consideration.

Initial features set

Since there were two different initial features sets being tested, 24 models had been built considering each of them (48 models variations ÷ 2 initial features sets). Their performance results (AUC and F_1-Score) were grouped and analyzed considering that partition, and table 6.11 was constructed. Additionally, this table also presents the results of the hypothesis tests performed to assess if the differences in performance, between the two initial features sets, were statistically significant.

Table 6.11: Summary of the models performance metrics, when grouping the results by the initial features set. Both the differences in AUC and F_1-Score, between the two initial features sets, are statistically significant, as demonstrated by the respective p-value.

Initial features set	AUC (mean ± SD)	AUC (median [first quartile / third quartile])	F_1-Score (mean ± SD)	F_1-Score (median [first quartile / third quartile])
'All'	0.91 ± 0.04	0.91 [0.88 / 0.93]	0.75 ± 0.09	0.78 [0.69 / 0.81]
'NoTemp&SpO_2'	0.86 ± 0.08	0.85 [0.81 / 0.93]	0.7 ± 0.1	0.76 [0.63 / 0.80]
p-value		< 0.01		< 0.01

Indeed, significant differences were found, which indicates that the initial features set 'All' consistently performed better than 'NoTemp&SpO_2'. Thereby, these results suggest that in a clinical context where it's possible to acquire BTemp and SpO_2 reliably, it might be advantageous to do so. Although, good performance in the deterioration prediction task can still be achieved without those vital signs information, as was concluded already when comparing the two best models.

Despite the significant differences, the fact that deterioration can still be accurately predicted using fewer sensors is highly desirable [10] and brings tremendous benefits. First, it promotes patient acceptability and enhances the ease of coping with continuous monitoring [10], while still assuring freedom of movement. Then, the fact that the SpO_2 sensor can be disregarded is convenient, since prior research reported that, despite usually accepted by patients, it is frequently removed and not returned to the finger [97]. Additionally, the SpO_2 sensor utilized in this thesis has revealed to be considerably unreliable and was the one that presented the shorter battery duration of the three.

These problems with the SpO_2 sensor highlight the fact that wearable sensors technology still presents some limitations at the moment (discussed in 1.1). These are particularly concerning when these sensors are to be employed outside of controlled clinical trials, which is yet not advised in this context [10]. Nonetheless, this technology has demonstrated significant progress over the last years [10] and studies like this one are important to validate their use.

In conclusion, this discussion and the results obtained, emphasize that finding the perfect balance between prediction performance and the number of wearable sensors used, may be critical for the success of continuous monitoring in becoming the standard practice for patient monitoring in general wards.

MEWS

Like what happened for the best models with initial features set 'All' and 'NoTemp&SpO_2', all the other 46 models variations attained far superior performance metrics than MEWS. This is highlighted by the fact that those models achieved an average AUC and F_1-Score of 0.88 ± 0.06 and 0.7 ± 0.1, respectively, and MEWS best configuration presented 0.76 AUC and 0.21 F_1-Score.

This overall better performance of ML models, when comparing with EWS, is consistent with previous research in this area [78], [103], [104], [116].

Since a discussion on this subject was already performed in 6.4.3.1 and the limitations of EWS were identified earlier in section 3.1, no further discussion is performed here.

6.5 Final warning system assembly

6.5.1 Methods

After identifying the best model, the final warning system was assembled. This was done with the purpose of estimating a realistic usage frequency for this system, in a real clinical context.

Since the best model was based on the LR algorithm, the regression coefficients were re-estimated using the entire dataset (considering both the training and test set of the corresponding dataset variation), as recommended in prior research [103], before the system assembly.

The system assembly itself consisted in connecting all the previously described development stages, as illustrated in figure 6.10. This way, when implemented in real time, and when presented with the arrival of a new 12-hours window of data from a certain patient, the system will (1) apply the respective physiological thresholding to the six continuous time series; (2) apply the acceptance criterion described in 6.1; (3) if the window fulfills step (2), apply the remaining preprocessing stages to the six continuous time series; (4) extract the features required by the final model (see 6.4.3.1); (5) rescale the numerical features and encode the categorical ones, as explained in 6.3; (6) impute this data to the final model and predict an outcome. This outcome is either 1, patient will deteriorate in the next 12 hours, or 0, patient won't deteriorate in the next 12 hours. The prediction result is then transmitted to the caregivers, as described in 4.1.1, and an alarm can be triggered in case the outcome reveals the possible occurrence of a deterioration event.

So that a realistic usage frequency could be estimated, the above-described process was simulated and timed using the final set of windows utilized for the development of prediction models. Only the steps in-between the brackets in figure 6.10 were timed, since these were the warning system's elements developed during this thesis. This measuring of the time each patient assessment takes, can be used to evaluate the system's practicality in a real context.

These simulations were conducted on a Lenovo Legion Y520 (CPU: Intel i5-7300HQ 2.50GHz, RAM: 8GB).

6.5.2 Results and discussion

The warning system took, on average, 47 ± 18 seconds[9] for each patient assessment. The maximum duration for an assessment was 121.5 seconds. None of the reviewed literature reported this result, hence no comparison can be performed.

In the light of these results, and by taking a pessimistic stand for the discussion, it can be assumed that the system will usually take around 2 minutes (the maximum duration) for the process in-between brackets in figure 6.10. Then, the two data transmission stages, outside the brackets, still have to be taken into consideration. These correspond to the

[9] median 44.5 (37.9/54.3) seconds (first quartile / third quartile)

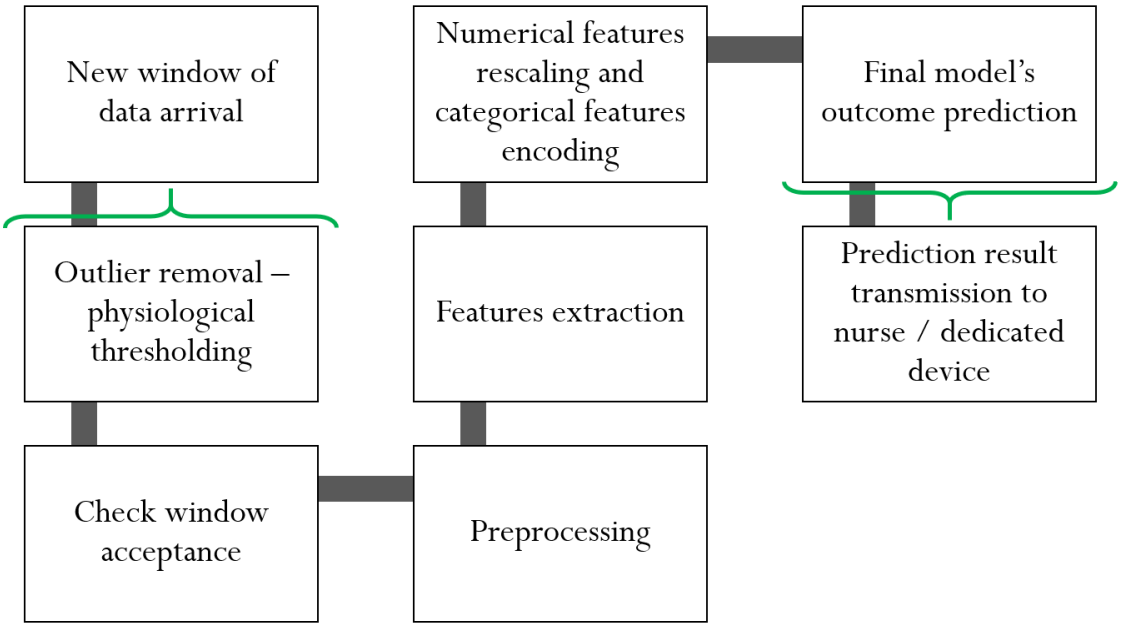


Figure 6.10: Final warning system workflow when implemented in real time and presented with a 12-hours window of data. 'Check window acceptance' refers to the acceptance criterion described in 6.1. 'Preprocessing' refers to the remaining preprocessing stages (excluding the already applied physiological thresholding). 'Numerical features rescaling and categorical features encoding' is performed as explained in 6.3. The elements in-between the green brackets correspond to the ones implemented during this thesis.

processes of transmitting the acquired data from the dedicated tablet to a server (see figure 4.2) and the prediction result transmission back to the dedicated tablet, where a nurse can have access to the result. Since these two processes are based on reliable wireless connections [122], such as wi-fi or cellular 3G/4G, they are expected to be considerably fast, of course, in the absence of transient problems within the hospital network infrastructure. Hence, there are reasons to believe that the warning system could seamlessly perform patients assessments every 10 minutes. However, even if a more pessimistic perspective is taken and patients assessments could only be conducted every 30 minutes or every hour, which is perfectly reasonable for the results attained, this would still mean that patient monitoring is being performed, at least, eight or four times more frequently than current practice. Besides the advantage of significantly shortening the time gaps without patient monitoring, which could lead to the first signs of deterioration to go unnoticed, this system has the additional convenience of not requiring nurses to interrupt their workflows to take vital signs measurements.

In summary, this warning system is fully independent of nurse's manual measurements of physiological signals, which enables a more frequent assessment of a patient's health status, without overloading nurses workflows. The system combines automatic continuous monitoring, provided by wearable sensors, with a ML-based decision support model. This automation of patient monitoring is imperative in general wards, where less

advanced equipment is present and where the monitoring conditions are poorer, and can contribute to improved patients outcomes.

Still, there is no replacement for a nurse's evaluation of the patient's condition. Therefore, these continuous monitoring-based warning systems should work as tools to help nurses gain insight on the patient's state and assist them in getting to the patient's side at the right time [114]. This way, a complete judgment on the patient's health status can be performed and a well-informed decision can be made regarding whether or not to put in motion therapeutic procedures.

Furthermore, this system relies on a LR model, which demonstrated remarkably superior performance when compared to a currently in-practice benchmark score, MEWS. This comprised a lower false alarm rate, which might considerably mitigate alarm fatigue in clinical staff, restoring their sensitivity towards alarms in wards. Also, patient's assessments that were followed by deterioration events were more accurately predicted. Despite these two improvements in performance can still be boosted to a greater extent, they already lead to smarter alarms, on which clinical staff can more confidently act upon. Additionally, important predictors and features were identified, which offers explainability and might lessen clinicians suspicion towards these ML-based decision support models. Not only that, this system explores correlations between vital signs and integrates demographic and contextual information, providing a more personalized patient monitoring. This is especially relevant for the monitoring of surgical patients, due to their particular characteristics [5].

6.6 Study limitations and future work

This study presented several limitations. First and foremost, this was a retrospective single-centered study, which included only 60 subjects from only two sub-populations of surgical patients (patients undergoing gastroesophageal cancer resection and patients undergoing hip fracture surgery), resulting in a very small dataset. Therefore, a larger and multicentered study, where more surgical patients sub-populations are considered, is required to validate the methodology employed, validate the results obtained and so that one could state that these results are generalizable. Only after that, and in case the outcomes are positive, should this warning system be adopted in a real clinical context. Secondly, the performance achieved in the prediction task, especially precision and recall, despite very promising, is probably not yet sufficient for practical implementation in a real hospital context. Thirdly, since this was a retrospective study, it could not be assessed how would clinical staff react to the fact that alarms were being generated by a ML-based decision support model, and how comfortable would they be with this novel situation. Lastly, several things could be changed/improved in the warning system development process. Particularly:

1. only one personalized approach to deal with missing data periods in vital signs time

series was developed, and only one other was described in more detail.

2. in particular for the personalized approach developed in this thesis, only one clustering algorithm was tested. Other algorithms, such as the ones listed in 2.4.2.3, could have been experimented as well.

3. for the prediction model development, only two ML algorithms were explored, since a completely comprehensive study on all ML algorithms would be unfeasible.

4. regarding the LR models development, a few considerations can be made:

    - predictors were constricted to be linearly combined. Instead, these could have been modeled with non-linear relationships, cubic splines and/or the inclusion of interaction terms.
    - rather than just displaying the discretized prediction outcome (0 or 1) to the nurses, the obtained probability itself could also be shown. This would allow clinical staff to be aware of the model's confidence in each particular prediction.
    - with a bigger dataset, confidence bounds surrounding the optimal probability threshold could be assertively identified. These would allow to define an "uncertainty" region, and different levels of alarms could be established. This is, probabilities below the lower bound would correspond to no alarm; probabilities between the confidence bounds would correspond to a level 1 alarm, where there's some likelihood of a patient deteriorating; probabilities above the upper bound would correspond to a level 2 alarm, where there's a high likelihood of a patient deteriorating.

5. the novelty detection approach was not explored, mostly due to time constrictions and because it requires high amounts of "Non-Event" data for the model of normality construction [97].

6. instead of only considering the current outcome of a prediction model to make a decision regarding the patient's condition, evaluating the outcome (e.g., the probability in LR) evolution in time could be performed. This strategy has already been discussed [5] and explored before [94], [103].

7. deterioration was tried to be predicted with 12 hours in advance and by analyzing the most recent 12 hours of patient data. Both these values, despite guided by prior research, might not be the optimal ones for the prediction task in this context. Thereby, different values could have been experimented, so that the optimal configuration could be found.

8. despite features importance in the final model was analyzed, the model does not state which physiological signal and/or feature deranged the most for each specific

warning. This might be an important add-on to guide clinicians on which health problems the patient might be undergoing.

9. features hyperparameters, such as the windows sizes used or α in feature B.1.5, could have been optimized.

Moreover, from the list of recommendations summarized in 3.2, regarding the development of prediction models, not all were considered. In particular, the inclusion of intra-operative predictors and the dynamic change of features relevance accordingly to the time-period of patient's stay. In a future study these should also be integrated.

Additionally, a couple of suggestions for a future study can be made. First, it is suggested, in this context, that studies report model's performance based on precision and recall, instead of AUC only. Second, the introduction of a new type of alarms can be proposed. Given that a wearable sensor can produce periods of missing or erroneous data, due to reasons such as sensor detachment or communication issues, a window acceptance criterion was implemented. This criterion was based on the percentage of samples that represent acceptable measures. Hence, it can be used as tool to control the sensors state. This is, if a predefined number of consecutive patients windows does not fulfill the acceptance criterion, clinical staff may receive an alarm advising to check the sensors placing and settings.

As final note, there were some reliability issues with the wearable sensors. The quality of the data was far from what was desired, which may have limited performance in the prediction task. Therefore, despite the promising results obtained, it is believed that further improvements are achievable. However, for that to happen, more studies exploring wearable sensors use are required to identify their limitations and understand how can these be overcome. This will contribute for this technology's progress, hopefully, up to a point where these sensors can be reliably utilized outside of controlled clinical trials.

Chapter 7

Conclusion

The limitations presented by monitoring systems currently in-use in general wards put surgical patients at risk of developing clinical deterioration events, during their ward stay. Those systems are based on EWS calculations and manual intermittent nurses controls, performed every 4 to 6 hours. This strategy may cause deterioration to remain unnoticed for hours, due to the strategy's periodical nature and to the inability of EWS to correctly predict the physiologically complex deterioration events experienced by these patients. This can lead to increased morbidity, mortality and severe deterioration events occurrence. Thus, a better clinical deterioration detection strategy was required.

With that in mind, in this thesis, it was hypothesized that by combining continuous vital signs monitoring, provided by wearable sensors, with ML-based prediction models, deterioration could be predicted earlier and more accurately, leading to better patients' outcomes.

Exploring that hypothesis, a warning system with those characteristics was developed. This system was fully independent of manual measurements, predicted deterioration, on average, with 7 ± 3 hours in advance and relied on a LR model that presented 85% recall, 79% precision, 98% specificity and 0.94 AUC. When comparing these results with MEWS, a commonly employed EWS, it was evident that MEWS was outperformed in every performance metric. A higher sensitivity and a lower false alarm rate, presented by the LR model developed, are particularly noteworthy, given the context of this work. Therefore, it can be stated that the main research goal was fulfilled, despite a study with a larger and multi-centered cohort being required to confirm these results.

For this goal's fulfillment, four more specific goals had to be satisfied (see 1.2). First, a thorough literature review was performed and a list of limitations and recommendations, regarding the warning system development, was produced (see 3.2) and a review table

with unique characteristics was constructed (see appendix A). In this thesis, all limitations were addressed and most recommendations were considered. Second, adequate preprocessing techniques were implemented for each type of data and two innovative solutions were developed. These were a new personalized approach to deal with periods of missing data in the vital signs time series and a novel variation of a RRI preprocessing technique for false beats correction. Third, a compilation of features, previously used in this context, was produced (see appendix B) and feature selection procedures were implemented to identify which features could actually provide insight about patterns of deterioration. Demographic and contextual information were found to be relevant and to significantly contribute for features deemed important in the deterioration prediction task. Also, RR and HR were found to be major predictors of deterioration, as in prior research. Lastly, a decision support model, based on LR, that can automatically warn clinicians in case of deterioration, was developed, as mentioned already.

No other study, in the reviewed literature, explored models that were simultaneously based on advanced algorithms to predict deterioration, and on continuous monitoring provided by wearable sensors alone (independent of manual measurements or bedside monitors). This highlights that more studies with these characteristics should be conducted and that this work marks an advance in the field, especially when the positive results obtained are considered.

This work's three main developments were the ML-based early warning system, a new personalized method to deal with periods of missing data in the vital signs time series and a novel variation of a RRI preprocessing technique for false beats correction.

Besides that, this study showed, once again, the EWS ineptitude for the deterioration prediction task and that these can easily be outperformed by ML-based prediction models.

Regarding the ML algorithms explored, LR was found to consistently perform better than BT. However, this was probably related with the small dataset size, and it cannot be assumed that LR will always perform better than BT. Both algorithms should be considered in a future study.

Regarding preprocessing techniques, it was found that personalized methods, to handle missing data periods in vital signs, contribute to significant improvements in prediction performance, when compared with simpler and generic approaches. In cases where only the latter are being considered, it was found that linear interpolation is the technique that more properly correct those gaps in the vital signs time series.

Regarding practical issues, it was concluded that the developed warning system can be employed, at least, eight or four times more frequently than current methods. Combining this information with the timely deterioration detection provided by the warning system (average 7 ± 3 hours in advance), this strategy's enhanced practicality and appropriateness are evidenced.

Additionally, it was demonstrated that deterioration can still be accurately predicted in the absence of BTemp and SpO_2 data. In fact, the final warning system did not require those sensors' information and depended on measures acquired from one sensor only,

which is extremely advantageous. Therefore, it was found that, in a clinical context where some sensors are unreliable, deterioration can still be feasibly detected using ML models.

In conclusion, this work provides support for wearable sensors to be employed, in combination with ML-based prediction models, for the automatic detection of clinical deterioration and endorses the implementation of continuous monitoring as standard practice for patients monitoring in wards.

Bibliography

[1] E. H. Gemmill, D. J. Humes, and J. A. Catton, "Systematic review of enhanced recovery after gastro-oesophageal cancer surgery," *Annals of the Royal College of Surgeons of England*, vol. 97, no. 3, pp. 173–179, 2015, ISSN: 00358843. DOI: 10.1308/003588414X14055925061630.

[2] M. Cardona-Morrell, M. Prgomet, R. M. Turner, M. Nicholson, and K. Hillman, "Effectiveness of continuous or intermittent vital signs monitoring in preventing adverse events on general wards: a systematic review and meta-analysis," *International Journal of Clinical Practice*, vol. 70, no. 10, pp. 806–824, 2016, ISSN: 17421241. DOI: 10.1111/ijcp.12846.

[3] M. Prgomet *et al.*, "Vital signs monitoring on general wards: Clinical staff perceptions of current practices and the planned introduction of continuous monitoring technology," *International Journal for Quality in Health Care*, vol. 28, no. 4, pp. 515–521, 2016, ISSN: 14643677. DOI: 10.1093/intqhc/mzw062.

[4] A. K. Khanna, P. Hoppe, and B. Saugel, "Automated continuous noninvasive ward monitoring: future directions and challenges," *Critical Care*, vol. 23, no. 1, p. 194, 2019, ISSN: 1364-8535. DOI: 10.1186/s13054-019-2485-7.

[5] C. Petit, R. Bezemer, and L. Atallah, "A review of recent advances in data analytics for post-operative patient deterioration detection," *Journal of Clinical Monitoring and Computing*, vol. 32, no. 3, pp. 391–402, 2018, ISSN: 15732614. DOI: 10.1007/s10877-017-0054-7.

[6] The Faculty of Intensive Care Medicine, "Guidelines for the provision of Intensive Care Services," Tech. Rep., 2019. [Online]. Available: https://www.ficm.ac.uk/standards-research-revalidation/guidelines-provision-intensive-care-services-v2.

[7] Royal College of Nursing (RCN) Policy Unit, *Policy Guidance 15/2006: Setting Appropriate Ward Nurse Staffing Levels in NHS Acute Trusts*, 2006. [Online]. Available: https://www.rcn.org.uk/about-us/our-influencing-work/policy-briefings/pol-1506.

[8] Royal College of Nursing Policy Unit, *Guidance on safe nurse staffing levels in the UK*, 2010. [Online]. Available: https://www.rcn.org.uk/professional-development/publications/pub-003860.

[9] The Centre for Clinical Practice at London: National Institute for Health and Clinical Excellence, *Acutely ill patients in hospital: Recognition of and response to acute illness in adults in hospital*. London, UK, 2007, pp. 1–107. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK45947/.

[10] C. Orphanidou, *Signal Quality Assessment in Physiological Monitoring: State of the Art and Practical Considerations*. Springer, 2018, ISBN: 978-3-319-68414-7. DOI: 10.1007/978-3-319-68415-4.

[11] M. A. DeVita *et al.*, ""Identifying the hospitalised patient in crisis"-A consensus conference on the afferent limb of Rapid Response Systems," *Resuscitation*, vol. 81, no. 4, pp. 375–382, 2010, ISSN: 03009572. DOI: 10.1016/j.resuscitation.2009.12.008.

[12] C. H. Leuvan and I. Mitchell, "Missed opportunities? An observational study of vital sign measurements.," *Critical care and resuscitation : journal of the Australasian Academy of Critical Care Medicine*, vol. 10, no. 2, pp. 111–115, 2008, ISSN: 14412772. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18522524/.

[13] J. Ludikhuize, S. M. Smorenburg, S. E. de Rooij, and E. de Jonge, "Identification of deteriorating patients on general wards; measurement of vital parameters and potential effectiveness of the Modified Early Warning Score," *Journal of Critical Care*, vol. 27, no. 4, 424.e7–424.e13, 2012, ISSN: 15578615. DOI: 10.1016/j.jcrc.2012.01.003.

[14] M. Odell, C. Victor, and D. Oliver, "Nurses' role in detecting deterioration in ward patients: Systematic literature review," *Journal of Advanced Nursing*, vol. 65, no. 10, pp. 1992–2006, 2009, ISSN: 03092402. DOI: 10.1111/j.1365-2648.2009.05109.x.

[15] R. M. Pearse *et al.*, "Mortality after surgery in Europe: A 7 day cohort study," *The Lancet*, vol. 380, no. 9847, pp. 1059–1065, 2012. DOI: 10.1016/S0140-6736(12)61148-9.

[16] M. Weenk *et al.*, "Wireless and continuous monitoring of vital signs in patients at the general ward," *Resuscitation*, vol. 136, pp. 47–53, 2019, ISSN: 18731570. DOI: 10.1016/j.resuscitation.2019.01.017.

[17] J. Wendon, C. Hodgson, and R. Bellomo, "Rapid response teams improve outcomes: we are not sure," *Intensive care medicine*, vol. 42, no. 4, pp. 599–601, 2016, ISSN: 14321238. DOI: 10.1007/s00134-016-4253-3.

[18] T. Kamio, A. Kajiwara, Y. Iizuka, J. Shiotsuka, and M. Sanui, "Frequency of vital sign measurement among intubated patients in the general ward and nurses' attitudes toward vital sign measurement," *Journal of Multidisciplinary Healthcare*, vol. 11, pp. 575–581, 2018, ISSN: 11782390. DOI: 10.2147/JMDH.S179033.

[19] H. Hogan *et al.*, "Preventable deaths due to problems in care in English acute hospitals: A retrospective case record review study," *BMJ Quality and Safety*, vol. 21, no. 9, pp. 737–745, 2012, ISSN: 20445415. DOI: 10.1136/bmjqs-2011-001159.

[20] A. H. Taenzer and B. C. Spence, "The Afferent Limb of Rapid Response Systems: Continuous Monitoring on General Care Units," *Critical Care Clinics*, vol. 34, no. 2, pp. 189–198, 2018, ISSN: 15578232. DOI: 10.1016/j.ccc.2017.12.001.

[21] Z. Sun *et al.*, "Postoperative Hypoxemia Is Common and Persistent," *Anesthesia & Analgesia*, vol. 121, no. 3, pp. 709–715, 2015, ISSN: 0003-2999. DOI: 10.1213/ANE.0000000000000836.

[22] Sensium, *Early detection of patient deterioration*. [Online]. Available: https://www.sensium.co.uk/ (visited on 09/21/2020).

[23] A. H. Taenzer, J. B. Pyke, S. P. McGrath, and D. S. Warner, "A review of current and emerging approaches to address failure-to-rescue," *Anesthesiology*, vol. 115, no. 2, pp. 421–431, 2011, ISSN: 15281175. DOI: 10.1097/ALN.0b013e318219d633.

[24] L. Goense *et al.*, "Hospital costs of complications after esophagectomy for cancer," *European Journal of Surgical Oncology*, vol. 43, no. 4, pp. 696–702, 2017, ISSN: 15322157. DOI: 10.1016/j.ejso.2016.11.013.

[25] P. H. Charlton, "Continuous respiratory rate monitoring to detect clinical deteriorations using wearable sensors," Ph.D. dissertation, King's College London, London, UK, 2017.

[26] M. D. Buist *et al.*, "Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care: A pilot study in a tertiary-care hospital," *Medical Journal of Australia*, vol. 171, no. 1, pp. 22–25, 1999, ISSN: 0025729X. DOI: 10.5694/j.1326-5377.1999.tb123492.x.

[27] M. Joshi *et al.*, "Wearable sensors to improve detection of patient deterioration," *Expert Review of Medical Devices*, vol. 16, no. 2, pp. 145–154, 2019, ISSN: 17452422. DOI: 10.1080/17434440.2019.1563480.

[28] I. J. Brekke, L. H. Puntervoll, P. B. Pedersen, J. Kellett, and M. Brabrand, "The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review," *PLOS ONE*, vol. 14, no. 1, 2019, ISSN: 19326203. DOI: 10.1371/journal.pone.0210875.

[29] E. Bose, L. Hoffman, and M. Hravnak, "Monitoring cardiorespiratory instability: Current approaches and implications for nursing practice," *Intensive and Critical Care Nursing*, vol. 34, pp. 12–19, 2016, ISSN: 09643397. DOI: 10.1016/j.iccn.2015.11.005.

[30] P. J. Watkinson *et al.*, "A randomised controlled trial of the effect of continuous electronic physiological monitoring on the adverse event rate in high risk medical and surgical patients," *Anaesthesia*, vol. 61, no. 11, pp. 1031–1039, 2006, ISSN: 00032409. DOI: 10.1111/j.1365-2044.2006.04818.x.

[31] C. P. Subbe, B. Duller, and R. Bellomo, "Effect of an automated notification system for deteriorating ward patients on clinical outcomes," *Critical Care*, vol. 21, no. 1, 2017, ISSN: 1466609X. DOI: 10.1186/s13054-017-1635-z.

[32] National Institute for Health and Clinical Excellence, *Visensia for early detection of deteriorating vital signs in adults in hospital: Medtech innovation briefing*, 2015. [Online]. Available: https://www.nice.org.uk/advice/mib36.

[33] M. R. Pinsky, G. Clermont, and M. Hravnak, "Predicting cardiorespiratory instability," *Critical Care*, vol. 20, no. 1, 2016, ISSN: 1364-8535. DOI: 10.1186/s13054-016-1223-7.

[34] B. Gross, D. Dahl, and L. Nielsen, "Physiologic monitoring alarm load on medical/surgical floors of a community hospital," *Biomedical Instrumentation and Technology*, vol. 45, no. s1, pp. 29–36, 2011, ISSN: 08998205. DOI: 10.2345/0899-8205-45.s1.29.

[35] E. P. Weledji and V. Verla, "Failure to rescue patients from early critical complications of oesophagogastric cancer surgery," *Annals of Medicine and Surgery*, vol. 7, pp. 34–41, 2016, ISSN: 20490801. DOI: 10.1016/j.amsu.2016.02.027.

[36] D. P. Raymond, *Complications of esophageal resection*. [Online]. Available: https://www.uptodate.com/contents/complications-of-esophageal-resection.

[37] M. Messager *et al.*, "Variations among 5 European countries for curative treatment of resectable oesophageal and gastric cancer: A survey from the EURECCA Upper GI Group (EUropean REgistration of Cancer CAre)," *European Journal of Surgical Oncology*, vol. 42, no. 1, pp. 116–122, 2016, ISSN: 15322157. DOI: 10.1016/j.ejso.2015.09.017.

[38] A. M. Almoudaris *et al.*, "Failure to rescue patients after reintervention in gastroesophageal cancer surgery in England," *JAMA Surgery*, vol. 148, no. 3, pp. 272–276, 2013, ISSN: 21686254. DOI: 10.1001/jamasurg.2013.791.

[39] L. A. Busweiler *et al.*, "Failure-to-rescue in patients undergoing surgery for esophageal or gastric cancer," *European Journal of Surgical Oncology*, vol. 43, no. 10, pp. 1962–1969, 2017, ISSN: 15322157. DOI: 10.1016/j.ejso.2017.07.005.

[40] E. C. Folbert *et al.*, "Complications during hospitalization and risk factors in elderly patients with hip fracture following integrated orthogeriatric treatment," *Archives of Orthopaedic and Trauma Surgery*, vol. 137, pp. 507–515, 2017, ISSN: 14343916. DOI: 10.1007/s00402-017-2646-6.

[41] T. Klestil *et al.*, "Impact of timing of surgery in elderly hip fracture patients: a systematic review and meta-analysis," *Scientific Reports*, vol. 8, 2018, ISSN: 20452322. DOI: 10.1038/s41598-018-32098-7.

[42] R. M. Padilla and A. M. Mayo, "Clinical deterioration: A concept analysis," *Journal of Clinical Nursing*, vol. 27, no. 7-8, pp. 1360–1368, 2018, ISSN: 13652702. DOI: https://doi.org/10.1111/jocn.14238.

[43] D. Jones, I. Mitchell, K. Hillman, and D. Story, "Defining clinical deterioration," *Resuscitation*, vol. 84, no. 8, pp. 1029–1034, 2013, ISSN: 03009572. DOI: 10.1016/j.resuscitation.2013.01.013.

[44] A. Coomarasamy *et al.*, "PROMISE: first-trimester progesterone therapy in women with a history of unexplained recurrent miscarriages – a randomised, double-blind, placebo-controlled, international multicentre trial and economic evaluation.," in *Health Technology Assessment, No. 20.41*, Southampton, UK, 2016, ch. Appendix 3. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK362736/.

[45] D. Dindo, N. Demartines, and P. A. Clavien, "Classification of surgical complications: A new proposal with evaluation in a cohort of 6336 patients and results of a survey," *Annals of Surgery*, vol. 240, no. 2, pp. 205–213, 2004, ISSN: 00034932. DOI: 10.1097/01.sla.0000133083.54934.ae.

[46] FDA, *What is a Serious Adverse Event?* [Online]. Available: https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event (visited on 01/12/2020).

[47] S. Moola, "Vital signs to monitor hospital patients: a systematic review," *JBI Library of Systematic Reviews*, vol. 6, no. 4, pp. 1–11, 2008, ISSN: 1838-2142. DOI: 10.11124/jbisrir-2008-785.

[48] The Heart Foundation, *What is normal blood pressure*. [Online]. Available: https://www.heartfoundation.org.au/heart-health-education/blood-pressure-and-your-heart (visited on 01/12/2020).

[49] CCM Health, *Body temperature - Definition*. [Online]. Available: https://health.ccm.net/faq/2498-body-temperature-definition (visited on 01/12/2020).

[50] C. R. Gomez, "Disorders of body temperature," in *Handbook of Clinical Neurology*, vol. 120, Elsevier B.V., 2014, pp. 947–957. DOI: 10.1016/B978-0-7020-4087-0.00062-0.

[51] J. L. Vincent *et al.*, "Improving detection of patient deterioration in the general hospital ward environmen," *European Journal of Anaesthesiology*, vol. 35, no. 5, pp. 325–333, 2018. DOI: 10.1097/EJA.0000000000000798.

[52] B. Hafen and S. Sharma, *Oxygen Saturation*. Treasure Island, Florida, USA: StatPearls Publishing, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK525974/.

[53] Healthline, *What Is a Normal Respiratory Rate for Kids and Adults?* [Online]. Available: https://www.healthline.com/health/normal-respiratory-rate (visited on 01/12/2020).

[54] M. A. Cretikos *et al.*, "Respiratory rate: The neglected vital sign," *Medical Journal of Australia*, vol. 188, no. 11, pp. 657–659, 2008, ISSN: 0025729X. DOI: 10.5694/j.1326-5377.2008.tb01825.x.

[55] C. Kelly, "Respiratory rate 1: why measurement and recording are crucial," *Nursing Times*, vol. 114, no. 4, pp. 23–24, 2018. [Online]. Available: https://www.nursingtimes.net/clinical-archive/respiratory-clinical-archive/respiratory-rate-1-why-measurement-and-recording-are-crucial-26-03-2018/.

[56] I. Wheatley, "Respiratory rate 3: how to take an accurate measurement," *Nursing Times*, vol. 114, no. 7, pp. 21–22, 2018. [Online]. Available: https://www.nursingtimes.net/clinical-archive/respiratory-clinical-archive/respiratory-rate-3-how-to-take-an-accurate-measurement-25-06-2018/.

[57] D. B. J. Tecelão, "Prediction of postoperative atrial fibrillation using the electrocardiogram: A proof of concept," M.S. thesis, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, Portugal, 2018.

[58] P. B. Oliva, S. C. Hammill, and W. D. Edwards, "Cardiac rupture, a clinically predictable complication of acute myocardial infarction: report of 70 cases with clinicopathologic correlations.," *Journal of the American College of Cardiology*, vol. 22, no. 3, pp. 720–726, 1993, ISSN: 0735-1097. DOI: 10.1016/0735-1097(93)90182-z.

[59] M. H. Vafaie, M. Ataei, and H. R. Koofigar, "Heart diseases prediction based on ECG signals' classification using a genetic-fuzzy system and dynamical model of ECG signals," *Biomedical Signal Processing and Control*, vol. 14, pp. 291–296, 2014, ISSN: 17468108. DOI: 10.1016/j.bspc.2014.08.010.

[60] H. Hakkak and M. Azarnoosh, "Analysis of lossless compression techniques time-frequency-based in ECG signal compression," *Asian Journal of Biomedical and Pharmaceutical Sciences*, vol. 9, no. 66, 2019. DOI: 10.35841/2249-622X.66.18-867.

[61] A. Szulewski, *Normal ECG*. [Online]. Available: https://elentra.healthsci.queensu.ca/assets/modules/ECG/normal_ecg.html (visited on 09/11/2020).

[62] ECG & Echo Learning, *ECG interpretation: Characteristics of the normal ECG (P-wave, QRS complex, ST segment, T-wave)*. [Online]. Available: https://ecgwaves.com/topic/ecg-normal-p-wave-qrs-complex-st-segment-t-wave-j-point/ (visited on 09/11/2020).

[63] McGrawHill, *QRS Complexes*, 2007. [Online]. Available: https://co.grand.co.us/DocumentCenter/View/639/QRS-Complexes-Fast-and-Easy-ECGs-Shade--Wesley.

[64] J. E. Madias, "Low QRS voltage and its causes," *Journal of Electrocardiology*, vol. 41, no. 6, pp. 498–500, 2008, ISSN: 00220736. DOI: 10.1016/j.jelectrocard.2008.06.021.

[65] L. Rosenthal, *Normal Electrocardiography (ECG) Intervals*. [Online]. Available: https://emedicine.medscape.com/article/2172196-overview (visited on 09/11/2020).

[66] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in Public Health*, vol. 5, p. 258, 2017, ISSN: 2296-2565. DOI: 10.3389/fpubh.2017.00258.

[67] L. Murukesan, M. Murugappan, M. Iqbal, and K. Saravanan, "Machine learning approach for sudden cardiac arrest prediction based on optimal heart rate variability features," *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 4, pp. 521–532, 2014, ISSN: 21567026. DOI: 10.1166/jmihi.2014.1287.

[68] J. Ramshur, "Design, Evaluation, and Application of Heart Rate Variability Analysis Software (HRVAS)," M.S. thesis, University of Memphis, Tennessee, USA, 2010. DOI: 10.13140/RG.2.2.33667.81444.

[69] A. I. Batchinsky *et al.*, "Rapid prediction of trauma patient survival by analysis of heart rate complexity: Impact of reducing data set size," *Shock*, vol. 32, no. 6, pp. 565–571, 2009, ISSN: 10732322. DOI: 10.1097/SHK.0b013e3181a993dc.

[70] M. A. Peltola, "Role of editing of R-R intervals in the analysis of heart rate variability," *Frontiers in Physiology*, vol. 3, p. 148, 2012, ISSN: 1664042X. DOI: 10.3389/fphys.2012.00148.

[71] Expert System, *What is Machine Learning? A definition*. [Online]. Available: https://expertsystem.com/machine-learning-definition/ (visited on 01/12/2020).

[72] T. M. Mitchell, *Machine Learning*. New York, New York, USA: McGraw-Hill Science/Engineering/Math, 1997, ISBN: 0070428077.

[73] Y. Baştanlar and M. Özuysal, "Introduction to machine learning," *Methods in Molecular Biology*, vol. 1107, pp. 105–128, 2014, ISSN: 10643745. DOI: 10.1007/978-1-62703-748-8_7.

[74] J. Brownlee, *Difference Between Classification and Regression in Machine Learning*. [Online]. Available: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/ (visited on 01/12/2020).

[75] J. Brownlee, *Introduction to Dimensionality Reduction for Machine Learning*. [Online]. Available: https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/ (visited on 09/14/2020).

[76] J. Brownlee, *Logistic Regression for Machine Learning*. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/ (visited on 09/15/2020).

[77] M. Szumilas, "Explaining odds ratios," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 19, no. 3, pp. 227–229, 2010, ISSN: 17198429. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/.

[78] M. M. Churpek *et al.*, "Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards," *Critical Care Medicine*, vol. 44, no. 2, pp. 368–374, 2016, ISSN: 15300293. DOI: 10.1097/CCM.0000000000001571.

[79] E. Fox and C. Guestrin. University of Washington Online, *Machine Learning*, 2020. [Online]. Available: https://www.coursera.org/specializations/machine-learning.

[80] M. Kearns, "Thoughts on hypothesis boosting," *Unpublished manuscript*, 1988. [Online]. Available: https://www.cis.upenn.edu/~mkearns/papers/boostnote.pdf.

[81] M. Kearns and L. G. Valiant, "Crytographic limitations on learning Boolean formulae and finite automata," in *Proceedings of the twenty-first annual ACM symposium on Theory of computing - STOC '89*, New York, New York, USA: Association for Computing Machinery, 1989, pp. 433–444, ISBN: 0897913078. DOI: 10.1145/73007.73049.

[82] R. E. Schapire, "The Strength of Weak Learnability," *Machine Learning*, vol. 5, pp. 197–227, 1990, ISSN: 15730565. DOI: 10.1023/A:1022648800760.

[83] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, New York, USA: Springer-Verlag New York, 2013, ISBN: 9781461468486. DOI: 10.1007/978-1-4614-6849-3.

[84] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000, ISSN: 0090-5364. DOI: 10.1214/aos/1016218223.

[85] T. Van Der Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints," *BMC Medical Research Methodology*, vol. 14, no. 137, 2014, ISSN: 14712288. DOI: 10.1186/1471-2288-14-137.

[86] J. Brownlee, *10 Clustering Algorithms With Python*. [Online]. Available: https://machinelearningmastery.com/clustering-algorithms-with-python/ (visited on 09/14/2020).

[87] J. Brownlee, *Supervised and Unsupervised Machine Learning Algorithms*. [Online]. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ (visited on 09/14/2020).

[88] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, pp. 283–304, 1998, ISSN: 13845810. DOI: 10.1023/A:1009769707641.

[89] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1997, pp. 21–34. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.9984.

[90] J. Brownlee, *What is the Difference Between Test and Validation Datasets?* [Online]. Available: https://machinelearningmastery.com/difference-test-validation-datasets/ (visited on 01/12/2020).

[91] C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," *QJM: An International Journal of Medicine*, vol. 94, no. 10, pp. 521–526, 2001, ISSN: 14602725. DOI: 10.1093/qjmed/94.10.521.

[92] M. J. Rothman, S. I. Rothman, and J. Beals, "Development and validation of a continuous measure of patient condition using the Electronic Medical Record," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 837–848, 2013, ISSN: 15320464. DOI: 10.1016/j.jbi.2013.06.011.

[93] H. Gao *et al.*, "Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward," *Intensive Care Medicine*, vol. 33, pp. 667–679, 2007, ISSN: 03424642. DOI: 10.1007/s00134-007-0532-3.

[94] M. A. Pimentel, D. A. Clifton, L. Clifton, P. J. Watkinson, and L. Tarassenko, "Modelling physiological deterioration in post-operative patient vital-sign data," *Medical and Biological Engineering and Computing*, vol. 51, no. 8, pp. 869–877, 2013, ISSN: 01400118. DOI: 10.1007/s11517-013-1059-0.

[95] M. A. F. Pimentel, D. A. Clifton, and L. Tarassenko, "Gaussian process clustering for the functional characterisation of vital-sign trajectories," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Southampton, UK, 2013, pp. 1–6, ISBN: 9781479911806. DOI: 10.1109/MLSP.2013.6661947.

[96] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko, "Identification of patient deterioration in vital-sign data using one-class support vector machines," in *Federated Conference on Computer Science and Information Systems*, Szczecin, Poland, 2011, pp. 125–131, ISBN: 9788360810224. [Online]. Available: https://ieeexplore.ieee.org/document/6078208.

[97] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 722–730, 2014, ISSN: 21682194. DOI: 10.1109/JBHI.2013.2293059.

[98] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 193–197, 2013, ISSN: 00189294. DOI: 10.1109/TBME.2012.2208459.

[99] S. Visweswaran *et al.*, "Learning patient-specific predictive models from clinical data," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 669–685, 2010, ISSN: 15320464. DOI: 10.1016/j.jbi.2010.04.009.

[100] A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar, "Personalized Risk Scoring for Critical Care Patients using Mixtures of Gaussian Process Experts," 2016. arXiv: 1605.00959.

[101] L. Tarassenko, A. Hann, and D. Young, "Integrated monitoring and analysis for early warning of patient deterioration," *British Journal of Anaesthesia*, vol. 97, no. 1, pp. 64–68, 2006, ISSN: 14716771. DOI: 10.1093/bja/ael113.

[102] S. Khalid, D. A. Clifton, L. Clifton, and L. Tarassenko, "A two-class approach to the detection of physiological deterioration in patient vital signs, with clinical label refinement," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1231–1238, 2012, ISSN: 10897771. DOI: 10.1109/TITB.2012.2212202.

[103] M. M. Churpek *et al.*, "Multicenter development and validation of a risk stratification tool for ward patients," *American Journal of Respiratory and Critical Care Medicine*, vol. 190, no. 6, pp. 649–655, 2014, ISSN: 15354970. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25089847/.

[104] G. J. Escobar *et al.*, "Early detection of impending physiologic deterioration among patients who are not in intensive care: Development of predictive models using data from an automated electronic medical record," *Journal of Hospital Medicine*, vol. 7, no. 5, pp. 388–395, 2012, ISSN: 15535592. DOI: 10.1002/jhm.1929.

[105] R. Pirracchio *et al.*, "Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study," *The Lancet Respiratory Medicine*, vol. 3, no. 1, pp. 42–52, 2015, ISSN: 22132619. DOI: 10.1016/S2213-2600(14)70239-5.

[106] F. Dal Canton, V. M. Quinten, and M. A. Wiering, "Early Detection of Sepsis Induced Deterioration Using Machine Learning," in *BNAIC 2018: Artificial Intelligence*, Springer, Cham, 2019, pp. 1–15, ISBN: 9783030319779. DOI: 10.1007/978-3-030-31978-6_1.

[107] N. Stevens *et al.*, "Smart alarms: Multivariate medical alarm integration for post CABG surgery patients," in *IHI '12: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, Miami, Florida, USA: Association for Computing Machinery, 2012, pp. 533–542, ISBN: 9781450307819. DOI: 10.1145/2110363.2110423.

[108] Y. Mao *et al.*, "Medical data mining for early deterioration warning in general hospital wards," in *Proceedings of the IEEE 11th International Conference on Data Mining Workshops*, Vancouver, British Columbia, Canada, 2011, pp. 1042–1049, ISBN: 9780769544090. DOI: 10.1109/ICDMW.2011.117.

[109] T. J. Moss *et al.*, "Cardiorespiratory dynamics measured from continuous ECG monitoring improves detection of deterioration in acute care patients: A retrospective cohort study," *PLOS ONE*, vol. 12, no. 8, pp. 1–16, 2017, ISSN: 19326203. DOI: 10.1371/journal.pone.0181448.

[110] M. Hravnak *et al.*, "Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system," *Archives of Internal Medicine*, vol. 168, no. 12, pp. 1300–1308, 2008, ISSN: 00039926. DOI: 10.1001/archinte.168.12.1300.

[111] M. Hravnak *et al.*, "Cardiorespiratory instability before and after implementing an integrated monitoring system," *Critical Care Medicine*, vol. 39, no. 1, pp. 65–72, 2011, ISSN: 15300293. DOI: 10.1097/CCM.0b013e3181fb7b1c.

[112] C. Downey, R. Randell, J. Brown, and D. G. Jayne, "Continuous versus intermittent vital signs monitoring using a wearable, wireless patch in patients admitted to surgical wards: Pilot cluster randomized controlled trial," *Journal of Medical Internet Research*, vol. 20, no. 12, 2018, ISSN: 14388871. DOI: 10.2196/10802.

[113] M. Helfand, V. Christensen, and J. Anderson, "Technology Assessment: EarlySense for Monitoring Vital Signs in Hospitalized Patients," VA ESP Project #09-199, Portland, Oregon, USA, Tech. Rep., 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK384615/pdf/Bookshelf_NBK384615.pdf.

[114] E. Zimlichman *et al.*, "Early recognition of acutely deteriorating patients in non-intensive care units: Assessment of an innovative monitoring technology," *Journal of Hospital Medicine*, vol. 7, no. 8, pp. 628–633, 2012, ISSN: 15535592. DOI: 10.1002/jhm.1963.

[115] J. S. Thompson *et al.*, "Temporal patterns of postoperative complications," *Archives of Surgery*, vol. 138, no. 6, pp. 596–603, 2003, ISSN: 00040010. DOI: 10.1001/archsurg.138.6.596.

[116] B. H. Cuthbertson, M. Boroujerdi, L. McKie, L. Aucott, and G. Prescott, "Can physiological variables and early warning scoring systems allow early recognition of the deteriorating surgical patient?" *Critical Care Medicine*, vol. 35, no. 2, pp. 402–409, 2007, ISSN: 00903493. DOI: 10.1097/01.CCM.0000254826.10520.87.

[117] J. F. Fieselmann, M. S. Hendryx, C. M. Helms, and D. S. Wakefield, "Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients," *Journal of General Internal Medicine*, vol. 8, no. 7, pp. 354–360, 1993, ISSN: 08848734. DOI: 10.1007/BF02600071.

[118] L. Chen *et al.*, "Using supervised machine learning to classify real alerts and artifact in online multisignal vital sign monitoring data," *Critical Care Medicine*, vol. 44, no. 7, pp. e456–e463, 2016, ISSN: 15300293. DOI: 10.1097/CCM.0000000000001660.

[119] J. Kellett and B. Deane, "The Simple Clinical Score predicts mortality for 30 days after admission to an acute medical unit," *QJM: An International Journal of Medicine*, vol. 99, no. 11, pp. 771–781, 2006, ISSN: 14602725. DOI: 10.1093/qjmed/hcl112.

[120] Isansys, *Wearable Sensors*. [Online]. Available: https://www.isansys.com/en/Wearable-Sensors (visited on 01/25/2020).

[121] Nonin, *WristOx2® Model 3150 OEM with Bluetooth® Low Energy*. [Online]. Available: https://www.nonin.com/products/3150-oem-ble/ (visited on 01/25/2020).

[122] Isansys, *Connectivity*. [Online]. Available: https://www.isansys.com/en/connectivity (visited on 01/25/2020).

[123] A. E. Johnson *et al.*, "Machine Learning and Decision Support in Critical Care," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444–466, 2016, ISSN: 15582256. DOI: 10.1109/JPROC.2015.2501978.

[124] L. H. Lehman, M. Saeed, G. B. Moody, and R. G. Mark, "Similarity-based searching in multi-parameter time series databases," in *Computers in Cardiology*, Bologna, Italy: IEEE, 2008, pp. 653–656, ISBN: 1424437067. DOI: 10.1109/CIC.2008.4749126.

[125] D. Sow, A. Biem, J. Sun, J. Hu, and S. Ebadollahi, "Real-time prognosis of ICU physiological data streams," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Buenos Aires, Argentina: IEEE, 2010, pp. 6785–6788, ISBN: 9781424441235. DOI: 10.1109/IEMBS.2010.5625983.

[126] S. Stuiver, M. Breteler, M. Hermans, H. Hermens, and C. Kalkman, "Continuous monitoring of thoracic skin and axillary temperature in high-risk surgical patients using wireless patch sensors - A Clinical Validation Study," *Unpublished manuscript*, 2019.

[127] J. Sun, D. Sow, J. Hu, and S. Ebadollahi, "A system for mining temporal physiological data streams for advanced prognostic decision support," in *Proceedings of the IEEE International Conference on Data Mining*, Sydney, New South Wales, Australia: IEEE, 2010, pp. 1061–1066, ISBN: 9780769542560. DOI: 10.1109/ICDM.2010.102.

[128] R. Dürichen, M. A. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, "Multitask Gaussian processes for multivariate physiological time-series analysis," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 314–322, 2015, ISSN: 15582531. DOI: 10.1109/TBME.2014.2351376.

[129] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013, ISSN: 00313203. DOI: 10.1016/j.patcog.2012.07.021.

[130] J. van den Hoven, "Clustering with optimised weights for Gower ' s metric," Ph.D. dissertation, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, 2016.

[131] B. C. Ross, "Mutual information between discrete and continuous data sets," *PLOS ONE*, vol. 9, no. 2, 2014, ISSN: 19326203. DOI: 10.1371/journal.pone.0087357.

[132] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066 138, 2004, ISSN: 1063651X. DOI: 10.1103/PhysRevE.69.066138.

[133] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, ISSN: 03770427. DOI: 10.1016/0377-0427(87)90125-7.

[134] L. Kaufman and P. J. Rousseeuw, *Finding groups in data : an introduction to cluster analysis*. New York, New York, USA: Wiley, 1990, ISBN: 9780471878766.

[135] D. J. Doyle, E. H. Garmon, A. Goyal, and P. Bansal, *American Society of Anesthesiologists Classification*. Treasure Island, Florida, USA: StatPearls Publishing, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK441940/.

[136] M. Altuve *et al.*, "Analysis of the QRS complex for apnea-bradycardia characterization in preterm infants," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, Minnesota, USA, 2009, pp. 946–949, ISBN: 9781424432967. DOI: 10.1109/IEMBS.2009.5333153.

[137] G. D. Clifford and L. Tarassenko, "Quantifying errors in spectral estimates of HRV due to beat replacement and resampling," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 4, pp. 630–638, 2005, ISSN: 00189294. DOI: 10.1109/TBME.2005.844028.

[138] E. Karey *et al.*, "The use of percent change in RR interval for data exclusion in analyzing 24-h time domain heart rate variability in rodents," *Frontiers in Physiology*, vol. 10, p. 693, 2019, ISSN: 1664042X. DOI: 10.3389/fphys.2019.00693.

[139] J. McNames, T. Thong, and M. Aboy, "Impulse rejection filter for artifact removal in spectral analysis of biomedical signals," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, California, USA, 2004, pp. 145–148, ISBN: 0780384393. DOI: 10.1109/iembs.2004.1403112.

[140] R. A. Thuraisingham, "Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR intervals," *Computer Methods and Programs in Biomedicine*, vol. 83, no. 1, pp. 78–82, 2006, ISSN: 01692607. DOI: 10.1016/j.cmpb.2006.05.002.

[141] P. S. Hamilton and W. J. Tompkins, "Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database," *IEEE Transactions on Biomedical Engineering*, vol. BME-33, no. 12, pp. 1157–1165, 1986, ISSN: 15582531. DOI: 10.1109/TBME.1986.325695.

[142] R. Logier, J. De Jonckheere, and A. Dassonneville, "An efficient algorithm for R-R intervals series filtering," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, California, USA, 2004, pp. 3937–3940, ISBN: 0780384393. DOI: 10.1109/iembs.2004.1404100.

[143] D. Morelli, A. Rossi, M. Cairo, and D. A. Clifton, "Analysis of the impact of interpolation methods of missing RR-intervals caused by motion artifacts on HRV features estimations," *Sensors*, vol. 19, no. 14, p. 3163, 2019, ISSN: 14248220. DOI: 10.3390/s19143163.

[144] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *International Journal of Advances in Soft Computing and its Applications*, vol. 7, no. 3, pp. 176–204, 2015, ISSN: 20748523. [Online]. Available: https://www.researchgate.net/publication/288228469.

[145] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, ISSN: 10414347. DOI: 10.1109/TKDE.2008.239.

[146] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," 2012, ISSN: 10495258. arXiv: 1206.2944.

[147] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," 2010. arXiv: 1012.2599.

[148] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0118432.

[149] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 233–240, ISBN: 1595933832. DOI: 10.1145/1143844.1143874.

[150] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014, ISSN: 15337928. [Online]. Available: https://dl.acm.org/doi/10.5555/2627435.2697065.

[151] P. C. Austin, D. S. Lee, E. W. Steyerberg, and J. V. Tu, "Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods?" *Biometrical Journal*, vol. 54, no. 5, pp. 657–673, 2012, ISSN: 03233847. DOI: 10.1002/bimj.201100251.

[152] D. A. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *Journal of Signal Processing Systems*, vol. 65, no. 3, pp. 371–389, 2011, ISSN: 19398018. DOI: 10.1007/s11265-010-0513-6.

[153] M. M. Churpek, R. Adhikari, and D. P. Edelson, "The value of vital sign trends for detecting clinical deterioration on the wards," *Resuscitation*, vol. 102, pp. 1–5, 2016, ISSN: 18731570. DOI: 10.1016/j.resuscitation.2016.02.005.

[154] Ó. D. Lara, A. J. Prez, M. A. Labrador, and J. D. Posada, "Centinela: A human activity recognition system based on acceleration and vital sign data," *Pervasive and Mobile Computing*, vol. 8, no. 5, pp. 717–729, 2012, ISSN: 15741192. DOI: 10.1016/j.pmcj.2011.06.004.

[155] M. Vollmer, "HRVTool - an Open-Source Matlab Toolbox for Analyzing Heart Rate Variability," in *2019 Computing in Cardiology Conference (CinC)*, vol. 46, Computing in Cardiology, 2019. DOI: 10.22489/cinc.2019.032.

[156] A. Jovic and N. Bogunovic, "Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features," *Artificial Intelligence in Medicine*, vol. 51, no. 3, pp. 175–186, 2011, ISSN: 09333657. DOI: 10.1016/j.artmed.2010.09.005.

[157] S. Pincus, "Approximate entropy (ApEn) as a complexity measure," *Chaos*, vol. 5, no. 1, pp. 110–117, 1995, ISSN: 10541500. DOI: 10.1063/1.166092.

[158] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Americal Journal of Physiology Heart and Circulatory Physiology*, vol. 278, no. 6, H2039–H2049, 2000. DOI: 10.1152/ajpheart.2000.278.6.H2039.

[159] V. Martínez-Cagigal, *Sample Entropy. Mathworks*, 2018. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/69381-sample-entropy (visited on 10/22/2020).

[160] D. S. Quintana and J. A. Heathers, "Considerations in the assessment of heart rate variability in biobehavioral research," *Frontiers in Psychology*, vol. 5, p. 805, 2014, ISSN: 16641078. DOI: 10.3389/fpsyg.2014.00805.

[161] M. K. Moridani, S. K. Setarehdan, A. Motie Nasrabadi, and E. Hajinasrollah, "Nonlinear feature extraction from HRV signal for mortality prediction of ICU cardiovascular patient," *Journal of Medical Engineering and Technology*, vol. 40, no. 3, pp. 87–98, 2016, ISSN: 1464522X. DOI: 10.3109/03091902.2016.1139201.

[162] A. Nait-Ali, *Biometrics Under Biomedical Considerations*. Singapore, Singapore: Springer Nature Singapore, 2019. [Online]. Available: https://doi.org/10.1007/978-981-13-1144-4.

Appendix A

Literature review table

This appendix includes a table that summarizes the reviewed work regarding new strategies for the development of clinical deterioration detection models.

Table A.1: Summary of the reviewed work regarding new strategies for the development of clinical deterioration detection models.

Study	Type of monitoring[a]	Outcomes included	Type of patient focus	Predictors	Vital signs analysis[b]	Personalization	Missing data approach	Data approach / Prediction strategy[c]	Classes ratio[d]	Train/test partition	Model approach	Results	Model observations
Tarassenko et al. [101] 2006	HR, RR, BTemp and SpO_2 - continuous (bedside monitors). BP - intermittent	NM	General ward (not specifically surgical patients)	Vital signs	Independently	None	Median value of the last 5 minutes or mean value in the training set	Every new set of measurements was stored as an observation for the model. Alarms were generated assessing 5 minutes windows	NM	Training - 150 patients from another previous study. Testing - 168 patients	Novelty detection	1 alarm every 7.8 hours on average. 95% of the generated alarms were true (assessed by clinicians). Alarms were generated earlier than single-channel alerts (based on one vital sign only)	Combination of k-means clustering and kernel density estimates. The data fusion method adopted in BioSign is a probabilistic model of normality in five dimensions, previously learned from the vital sign data acquired from a representative sample of patients
Kellett et al. [119] 2006[e]	Used information obtained at time of admission	Death within 30 days of hospital admission	Acute care settings patients (higher monitoring than general wards)	Demographics, vital signs, functional status, prior health status and social measures	Independently	Included demographic, social and prior health status features	Excluded patients with missing data	NA	Training cohort: 316:6420 = 1:20.32	68%/32%	Logistic regression to identify predictors	Mortality prediction within 24 hours: AUC 0.91. Mortality prediction within 30 days: AUC 0.86	LR identified 16 independent predictors of 30 day mortality, from which the Simple Clinical Score was derived, stratifying patients into five risk classes
Cuthbertson et al. [116] 2007	Intermittent	ICU transfer	High dependency unit patients (higher monitoring than general wards)	Vital signs	Independently	None	NM	For patients that deteriorated: data from the last 48 hours before ICU admission. For the others: all available data until discharge	Patient's ratio: 67:69 ≈1:1	NM. The use of a validation set was also not mentioned, which means results might be biased and require further validation	Logistic regression	HR: sensitivity 0.67, specificity 0.77, AUC 0.74. RR: sensitivity 0.70, specificity 0.86, AUC 0.82. SpO_2: sensitivity 0.66, specificity 0.86 AUC 0.79. f2[f]: sensitivity 0.77, specificity 0.88, AUC 0.88. SpO_2 and f2 could detect differences 48 hours before ICU admission with 0.79 AUC; 24 hours with 0.84 and 0.88; 12 hours with 0.82 and 0.89. (EWS had 0.79, 0.79 and 0.81)	Developed several models: some based on one vital sign only and others based on different vital signs combinations

Visweswaran et al. [99] 2010 (heart failure dataset)	Used information obtained at time of admission	Death and other serious complications	Not clear	Demographics, contextual data, vital signs, laboratory tests and ECG and radiographic findings	Independently	Several personalized steps in the developed algorithm	Iterative non-parametric algorithm described in the paper	NA	Patient's ratio - 1281:11178 = 1:8.73	66.66%/33.33%	Markov blankets with Bayesian model averaging	Some performance measures were better with statistical significance than a LR model (AUCs surrounding 0.60)	Computationally expensive. The patient-specific algorithm uses Markov blanket, Bayesian model averaging over a set of models and employs a patient-specific heuristic to locate a set of suitable models to average over
Clifton et al. [96] 2011	Intermittent	NM	Postoperative UGI step-down unit patients (higher monitoring than general wards)	Vital signs	Independently	None	Excluded observations with missing data	Clinicians reviews' defined each set of measurements as "abnormal" or "normal"	130:1370 = 1:10.50 (however, only normal data was used for training)	Training - 1240 patients. Validation - 130 patients. Testing - 130 patients. This was repeated 50 times. Mentioned the use of cross-validation in the training set for parameters optimization	Novelty detection	Best model: One-class SVM, accuracy 0.95, sensitivity 0.98, specificity 0.92	Gaussian mixture model and a one-class SVM
Mao et al. [108] 2011	Data obtained from the Electronic Medical Record (intermittent)	ICU transfer for training and also death for testing	General ward (not specifically surgical patients)	Demographics, vital signs and laboratory tests	Independently	Used demographic features	Last value and mean value of the patient's observations	Used the previous 24 hours of data and divided it into six 4 hours buckets	Patient's ratio - 1295:28927 = 1:22.30. Used a novel bagging method and exploratory undersampling to address class imbalance	Trained in the entire data and tested in a different dataset in real time	Logistic regression, SVM and decision tree	Transfer to ICU was predicted with sensitivity of 0.41, specificity of 0.95, AUC 0.80, precision 0.30, NPV 0.96, accuracy 0.92. Predicted ICU transfer with 4 hours in advance. Predicted death with 30 hours in advance. Alerts identified 55% of patients who died. Alerts identified 42% of patients who were transferred to an ICU	Threshold decision in the LR model was defined through the ROC curve for a specificity of 95%

Khalid et al. [102] 2012	Continuous (bedside monitors)	Outcomes that would cause the vital signs to surpass the Rapid Response Teams calling criteria	Step-down unit patients (higher monitoring than general wards)	Vital signs	Independently	None	Last value	Experts labeled periods of at least 4 minutes that exceeded Rapid Response Teams calling criteria as "normal" or "abnormal"	1215:1215 = 1:1	75%/25%. This was repeated 50 times	Multilayer perceptron, SVM and gaussian processes	Best model: Two-class SVM with novel label refinement, FP 5.00%, FN 5.25%	Combination of k-means clustering and kernel density estimates to refine labels
Zimlichman et al. [114] 2012 (EarlySense evaluation)	HR and RR - continuous (bedside monitors)	ICU transfer, patient intubation and mechanical ventilation on the ward and cardiac arrest while in the unit	General ward (not specifically surgical patients)	HR and RR	Analyzed combined thresholds	None	NM	For threshold alerts - analyzed the data every time measurements were updated. For trend alerts - analyzed the previous 24 hours of measurements. If any of these was followed by a major clinical event within 24 hours, it was considered a true case for alarm	Patient's ratio - 9:104 = 1:11.56	Mentioned a partition but did not specify	Univariate EWS-like thresholds and 24 hours trends thresholds	Threshold alerts - HR: sensitivity 0.82, specificity 0.67, AUC 0.74 and precision 0.21. RR: sensitivity 0.64, specificity 0.81, AUC 0.69 and precision 0.26. HR and RR: sensitivity 0.55, specificity 0.94, AUC 0.75 and precision 0.50. Trend alerts - HR: sensitivity 0.78, specificity 0.90, AUC 0.90 and precision 0.41. RR: sensitivity 1.00, specificity 0.64, AUC 0.85 and precision 0.20. HR and RR: sensitivity 0.78, specificity 0.94, AUC 0.93 and precision 0.54	Optimal thresholds were chosen to yield the maximal sum of sensitivity and specificity
Stevens et al. [107] 2012[e]	Continuous (bedside monitors)	All that could action an alarm	ICU (post coronary artery bypass graft surgery patients)	Vital signs	Independently	Included demographic, contextual and medical history features	NM	NA (focus was on comparing the amount of generated alarms with other simpler methods)	NM	NA	Fuzzy logic classifiers	Decreased clinical alarms by an average of 59% ± 17%	The fuzzy rules and thresholds were constructed through interviews with doctors and nurses

Escobar et al. [104] 2012	Data obtained from the Electronic Medical Record (intermittent)	ICU transfer and death	General Ward (wide variety of patient's populations)	Vital signs, laboratory tests, severity of illness scores, longitudinal chronic illness burden scores, hospital length of stay, and care directives	Independently	Used demographic features (such as age and sex)	Imputed normal values or dropped windows with missing data	Discrete time analysis in 24 hours (to predict in the following 12 hours based on the previous 24 hours)	4036:39782 = 1:9.86	50%/50%	Logistic regression	AUC 0.78 when including all groups of patients. To identify 44% of ICU transfers emitted 34 false alarms (MEWS emitted 69). Predicts with 12 hours in advance	Employed a technique to determine if any predictor-outcome relationship was nonlinear. Implemented a solution to deal with predictors with U-shaped risk
Pimentel et al. [94] 2013	Intermittent	ICU transfer and in-hospital death	Postoperative UGI general ward patients	Vital signs	Independently	None	Excluded observations with missing data	Used windows from the day of discharge of normal patients to train the model. Measured the distance between patient's condition on five different occasions during full hospital stay to prove that are differences between admission and discharge	Patient's ratio - 17:154 = 1:9.05	Trained with windows from the day of discharge of normal patients	Novelty detection	Significant differences were found between the novelty score for "normal" patients and for "abnormal" patients (one example was 12 hours before)	Kernel density estimates and implemented three different distance metrics
Pimentel et al. [95] 2013	Intermittent	ICU transfer and in-hospital death	Postoperative UGI general ward patients	HR and RR	Independently	None	Excluded observations with missing data	Used the last available 48 hours segments of data per patient	Patient's ratio - 16:138 = 1:8.60	65% of the "normal" patients/35% of the "normal" patients and 16 "abnormal" patients. This was repeated 50 times	Novelty detection	Sensitivity 0.68, specificity 0.66 and AUC 0.69	Gaussian process regression to analyze distances between vital signs trajectories of "normal" and "abnormal" patients

Clifton et al. [98] 2013	HR and SpO_2 - continuous (wearable sensors). BP and RR - intermittent	ICU transfer and death	Postoperative UGI general ward patients	Vital signs	Independently	Personalized gaussian process framework for handling periods of data uncertainty	Using the personalized gaussian process framework	Used a strategy described elsewhere [152]	NM	75%/25% for each patient dataset. Mentioned the use of cross-validation in the training set for parameters optimization	Novelty detection	Showed that with a personalized approach, clinical deterioration alerts are generated earlier	None
Clifton et al. [97] 2014	HR and SpO_2 - continuous (wearable sensors). BP and RR - intermittent	ICU transfer and death	Postoperative UGI general ward patients	Vital signs	Independently	None	Mean value of the patient's observations	Clinicians' reviews defined 1 hour intervals as "normal" or "abnormal" (true label). If a novelty threshold was exceed for 4 or more minutes in any 5 minutes window the model would consider that 1 hour interval as a positive case of deterioration	Patient's ratio - 37:163 = 1:4.40	Training - 126 patients. Validation - 36 patients. Testing - 38 patients. This was repeated 50 times	Novelty detection	Best model: One-class SVM, accuracy 0.94, sensitivity 0.96, specificity 0.93. Predictions are being made with 1 hour or less in advance	One-class SVM, one-class gaussian process, gaussian mixture model and kernel density estimate

Churpek et al. [103] 2014	Data obtained from the Electronic Medical Record (intermittent)	Cardiac arrest, ICU transfer and death	General ward (not specifically surgical patients)	Demographics, vital signs, mental status measures and laboratory tests	Tested interaction terms in the LR model	Used demographic features (such as age)	Last value and median in the patient's observations	Discrete time analysis in 8 hours intervals (to predict in different time spans and different outcomes based on the previous 8 hours of data)	Patient's ratio - 16452:269999 = 1:16.40	Training in the first 60% of the data (with 10-fold cross validation for model tuning) and testing in the remaining 40%	Logistic regression	This model was more accurate than the MEWS for detecting all outcomes using whether an event occurred within 24 hours of an observation - Cardiac arrest: AUC 0.83, ICU transfer: AUC 0.75, death: AUC 0.93, combined outcome: AUC 0.77. At a specificity of 90%, the model had a sensitivity of 54% for cardiac arrest within 24 hours compared with 39% for the MEWS. At a similar sensitivity (65%) the model had a specificity of 85% vs 71% for the MEWS	Interaction terms did not improve model performance so they were removed. After model validation, regression coefficients for the final model were re-estimated using the entire dataset. The predictors were discretized by using cut-off points
Pirracchio et al.[105] 2015	Intermittent	In-hospital death	ICU	Demographics, vital signs, laboratory tests and other contextual and clinical data	Tested interaction terms in the LR model	None	NM	Not clear	Patient's ratio - 3002:24508 = 1:8.16	10-fold cross-validation and additional validation in a smaller external dataset	Combination of ML algorithms	Best model: Super Learner, AUC 0.88 (0.94 in the external dataset) and RF, AUC 0.88	Some of the ML algorithms tested: LR, tree-based models, NN and boosted models
Churpek et al. [78] 2016	Data obtained from the Electronic Medical Record (intermittent)	Cardiac arrest, ICU transfer and death	General ward (not specifically surgical patients)	Demographics, contextual data, vital signs and laboratory tests	Investigated relations between vital signs after model development (didn't use this information to produce the model)	Used demographic features (such as age)	Last value and median in the patient's observations	Discrete time analysis in 8 hours intervals (to predict in the following 8 hours based on the previous 8 hours)	10309:10309 = 1:1	Training in the first 60% of the data (with 10-fold cross validation for model tuning) and testing in the remaining 40%	Tested several ML algorithms	Best model for combined outcome: RF - AUC 0.80. LR with linear - AUC 0.74. LR with spline - AUC 0.77. MEWS - AUC 0.70. Predicts with 8 hours in advance	Required a lot of information from the Electronic Medical record. ML algorithms tested: Decision trees, bagged trees, RF, boosted trees, kNN, NN, SVM, LR with linear predictors and LR with restricted cubic splines

Alaa et al. [100] 2016	Data obtained from the Electronic Medical Record (intermittent)	ICU transfer	General ward (not specifically surgical patients)	Vital signs and laboratory tests	Independently	Several personalized steps, which include the use of demographic, contextual and other features obtained at time of admission	NM	Utilized all patient's data to yield a time-evolving risk score and each patient was labeled as stable or deteriorating	Patient's ratio - 525:5788 = 1:11.02	10-fold stratified cross validation	Gaussian processes to learn classes of clinically deteriorating and stable patients	Outperformed MEWS, Rothman Index and a LR model in terms of sensitivity and precision	The algorithm aims to estimate the number of latent classes in the patients' population, due to its heterogeneity, and trains a mixture of gaussian process experts, where each expert models the physiological data streams associated with a specific class. The weights of each expert for each patient are personalized and calculated using features obtained at time of admission
Chen et al. [118] 2016[e]	HR, RR and SpO_2 - continuous (bedside monitors and sensors). BP - intermittent	None. This study focused on classifying threshold-based alerts as real or artifacts	Step-down unit patients (higher monitoring than general wards)	Vital signs	Independently	None	Excluded features with missing values when used in algorithms that cannot handle missing data, while keeping the full set of features for algorithms that can handle missingness	Used VSAE (vital signs alert events) epochs. VSAE epochs lasted at least 3 minutes and with a duty cycle of 2/3 (at least six of nine consecutive values over specified thresholds)	Train - 418 real, 158 artifact (1:2.65). Test - 327 real, 70 artifact (1:4.67)	Train - 576 VSAE epochs annotated by experts and used. 10 fold cross validation in training set to identify best models. Test - 397 VSAE epochs annotated by experts and used	Tested several ML algorithms	Best model (all using RF): SpO_2, AUC 0.79 to 0.87. BP, AUC 0.77 to 0.87. RR, AUC 0.85 to 0.97. HR, NA	Univariate alert type but used multivariate features to build the models for each vital sign. ML algorithms tested: kNN (at various k), Naïve Bayesian classifier, LR, SVM, and RF. An expert committee had previously annotated the alerts
Moss et al. [109] 2017	Vital signs and laboratory tests - intermittent. ECG - continuous (bedside monitors)	ICU transfer and unexpected death	Acute care settings patients (higher monitoring than general wards)	Vital signs, laboratory tests and ECG-based measures	Independently	None	Last value and median values for the vital signs and laboratory tests	Data from the previous 24 hours	23881:2194077 = 1:91.88	Bootstrap resampling	Logistic regression	Only with ECG: AUC 0.65. Only laboratory tests: AUC 0.63. Only vital signs: AUC 0.69. Laboratory tests and ECG: AUC 0.70. Vital signs and ECG: AUC 0.70. Vital signs and laboratory tests: 0.71. All: AUC 0.73. Prediction within the next 24 hours	LR predictors modeled with restricted cubic splines

DalCanton et al. [106] 2018	Continuous during the first 48h after admission (bedside monitors)	ICU transfer, death, kidney failure, liver failure and respiratory failure	Emergency department	ECG, RR and SpO_2	Independently	None	None used	Feature extraction method 1: 5 minutes long windows of data. Feature extraction method 2: 10 minutes long windows of data. Feature extraction method 3: sets of 10000 heartbeats. Assigned the labels to the windows/sets accordingly to the patient's label	Dependent on the feature extraction method but was always around 1:1	10-fold cross-validation with additional cross-validation for parameters tuning	Tested several ML algorithms	Best model: Two-class SVM, accuracy 0.62	ML algorithms tested: LR, Gradient Boosting Machines, RF, SVM, Multilayer Perceptrons, Naïve Bayes Classifers, and kNN

Abbreviations: HR - Heart Rate, RR - Respiration Rate, BTemp - Body Temperature, SpO_2 - (Peripheral) Oxygen Saturation, BP - Blood Pressure, NM - not mentioned, NA - Not applicable, LR - Logistic Regression, ICU - Intensive Care Unit, AUC - Area Under the receiver operating characteristic Curve, EWS - Early Warning Score, UGI - Upper gastrointestinal, SVM - Support Vector Machine, NPV - Negative Predictive Value, ROC - receiver operating characteristic, FP - False Positives, FN - False Negatives, MEWS - Modified Early Warning Score, ML - Machine Learning, RF - Random Forest, NN - Neural Networks, kNN - k-Nearest Neighbors, ECG - Electrocardiogram.

[a] Type of monitoring - whenever "intermittent", it is implicit that data is acquired manually

[b] Vital signs analysis - if independently or if correlations between them were assessed.

[c] Data approach / Prediction strategy - which data periods the researchers used and/or how did they label it / organize it and/or the approach employed to say that deterioration is being predicted.

[d] Classes ratio - ratio between the number of observations labeled as "abnormal" and "normal". When instead is the patient's ratio being displayed in the table, it means that the classes ratio wasn't mentioned in the paper.

[e] different study purpose.

[f] f2 - discriminant function that combined HR, RR and SpO_2.

Appendix B

List of features

This appendix includes the list of features implemented. Some of them were adapted from previous studies and others are novel features designed during this thesis. Unless specified otherwise, the features were extracted considering the entire 12-hours window.

B.1 Numerical features

B.1.1 RR-SpO_2 ratio

These features' development was prompted by a finding made by Tarassenko et al. [101]. They reported that sudden deterioration cases were often preceded by elevated RR and by a gradual decrease in SpO_2. For that reason, two features that try to capture this evolution were designed.

The first, $full\_ratio\_rrsp$, is simply the ratio between the mean RR and the mean SpO_2 in the window.

The second, *slope_rrsp*, is obtained by first dividing the 12-hours window into four 3-hours windows. Then, the above-mentioned ratio is calculated for each of the smaller windows and a linear regression is fitted to those results. Finally, *slope_rrsp* corresponds to the slope of the fitted line.

B.1.2 Vital signs differences from normality

This is a personalized feature that is calculated separately for each vital sign. It intends to compare current vital sign's values with a normal and healthy one, through the calculation of:

$$V\_normdiff = \frac{|V - V_{normal}|}{V_{normal}} \tag{B.1}$$

where $V\_normdiff$ is the resulting feature and V is the vital sign mean in the window's most recent 3 hours of data. V_{normal} can be determined in two different ways. If the subject had a preoperative baseline measure available for this vital sign, this would be V_{normal}. Otherwise, the 10 closest "Non-Event" subjects would be identified (like in the new approach, see figure 5.1, but retrieving only "Non-Event" subjects) and their most recent 3 hours of data was averaged. Then, V_{normal} would be the average over that set of values.

In the first case, the preoperative baseline measure was used because, at least for elective surgery patients, this can be considered a normal value for this patient's vital sign [5]. In the second case, since only patients demographically similar to the patient being considered, and that did not deteriorate, are being utilized, this can be an adequate estimation for this patient's normal vital sign value. The most recent data from the closest patients is the one being averaged because this is more likely to represent a stable period. The reason for this is that this is the moment furthest away from surgery and closer to the time of hospital discharge.

B.1.3 RR samples above 27 breaths/minute

This feature was inspired by a result obtained by Fieselmann et al. [117], where it was reported that the occurrence of multiple RR observations above 27 breaths/minute, over a 72 hours period, was a rule with promising results in predicting cardiopulmonary arrest.

In this thesis, however, what was calculated was the percentage of available samples that were above the 27 breaths/minute threshold.

B.1.4 Trends

This feature extraction procedure was applied to the four vital signs and to the QRSa time series. It intends to obtain an estimation of the overall trend of each of those signals, across the 12-hours window.

First, a detrended time series was calculated using first-order polynomial detrending. Then, by subtracting it to the original one, the trend time series can be obtained. This would always be a line, since first-order detrending was utilized. Finally, the slope of this line was determined, which would correspond to the feature.

Additionally, this was also performed by fitting a robust linear regression [118] to the time series. The respective slope would correspond to the feature.

At last, the trends of the most recent 15 and 60 minutes windows [124] were also calculated by fitting linear regressions to the data. The respective slopes and intercept terms were kept as features.

B.1.5 Exponential smoothing average

This feature was extracted from the four vital signs time series. For its implementation, the 12-hours window was first divided into four 3-hours windows. Then, the following was calculated [153]:

$$S_t = \begin{cases} x_t, & t = 1 \\ \alpha x_t + (1-\alpha) S_{t-1}, & t = 2, 3, 4 \end{cases} \tag{B.2}$$

where S_t is the smoothing average for window t, x_t is the average over all samples in window t and α is a smoothing factor that allows to control the weight that is given to the current and previous windows. $\alpha = 0.5$ was used for all vital signs.

The feature itself corresponds to S_4. Hence, this is a weighed measure of the average vital sign value in the 12-hours window, that assigns a higher weight to more recent measures.

B.1.6 Vital signs frequency domain features

Two frequency domain features, adapted from Chen et al. [118], were extracted from the vital signs time series. Since these time series could contain missing values, the technique employed for power spectrum estimation should be able to handle them. For that reason, the Lomb-Scargle periodogram was used.

The first feature, *spectrum_ratio*, was the ratio between high frequencies power and low frequencies power. Frequencies above half the maximum frequency present in the spectrum were considered high frequencies. The remaining were considered low frequencies.

The second feature, $maxpow\_freq$, corresponds to the frequency that presented the maximum power.

B.1.7 Basic statistical features

Several basic statistical measures were extracted. Table B.1 summarizes them, providing yet a short feature description, the information regarding the portion of the 12 hours window that was used and the information regarding which signals the respective feature was extracted from.

Table B.1: Summary of the basic statistical features extracted.

Feature	Description	Window portion(s) considered	Signals
Average	Average over all available samples	12 hours	Vital signs, QRSa and RRI
Recent average	Average over all available samples	Most recent 3 hours (30 minutes for RRI [109])	Vital signs, QRSa and RRI

Last 5 average	Average over all available samples	Most recent 5 minutes [136]	QRSa
Median	Median over all available samples	12 hours	Vital signs, QRSa and RRI
Recent median	Median over all available samples	Most recent 3 hours	Vital signs and QRSa
Standard deviation	Standard deviation over all available samples	12 hours	Vital signs, QRSa and RRI
Recent standard deviation	Standard deviation over all available samples	Most recent 3 hours (30 minutes for RRI [109])	Vital signs, QRSa and RRI
Last 5 standard deviation	Standard deviation over all available samples	Most recent 5 minutes [136]	QRSa
Minimum	Minimum value amongst all available samples	12 hours	RRI, HR, RR and SpO_2*
Recent minimum	Minimum value amongst all available samples	Most recent 3 hours	HR, RR and SpO_2*
Maximum	Maximum value amongst all available samples	12 hours	Vital signs and RRI
Recent maximum	Maximum value amongst all available samples	Most recent 3 hours	Vital signs
Worst	The most deranged value (the furthest away from healthy values)	12 hours	HR and RR**
Recent worst	The most deranged value (the furthest away from healthy values)	Most recent 3 hours	HR and RR**
Range	Subtraction between maximum and minimum [104]	12 hours	HR, RR and SpO_2*
Recent range	Subtraction between maximum and minimum [104]	Most recent 3 hours	HR, RR and SpO_2*
Coefficient of variation	$\frac{Standard\_deviation}{Average}$ [118]	12 hours	Vital signs and QRSa
Recent coefficient of variation	$\frac{Recent\_standard\_deviation}{Recent\_average}$ [118]	Most recent 3 hours	Vital signs and QRSa
Range ratio	$\frac{Range}{Median}$ [118]	12 hours	HR, RR and SpO_2*
Recent range ratio	$\frac{Recent\_range}{Recent\_median}$ [118]	Most recent 3 hours	HR, RR and SpO_2*

Median of absolute deviation from the median	$med\{\|x_i - Median\|\}$ [118]	12 hours	Vital signs and QRSa
Recent median of absolute deviation from the median	$med\{\|x_i - Recent\_median\|\}$ [118]	Most recent 3 hours	Vital signs and QRSa
Delta average	$\|Recent\_average - Initial\_average\|$	Most recent 3 hours (for $Recent\_average$) and the initial 3 hours of the window (for $Initial\_average$)	Vital signs, QRSa and RRI
Delta median	$\|Recent\_median - Initial\_median\|$	Most recent 3 hours (for $Recent\_median$) and the initial 3 hours of the window (for $Initial\_median$)	Vital signs, QRSa and RRI
Delta standard deviation	$\|Recent\_std - Initial\_std\|$ (std - standard deviation)	Most recent 3 hours (for $Recent\_std$) and the initial 3 hours of the window (for $Initial\_std$)	Vital signs and QRSa

Abbreviations: HR - Heart rate, RR - Respiration Rate, BTemp - Body Temperature, SpO_2 - (Peripheral) Oxygen Saturation, QRSa - QRS complex amplitude, RRI - RR interval.
$med\{\cdot\}$ is the median operator.
* BTemp was not considered due to the sensor issue described in 5.1.1.
** BTemp was not considered due to the sensor issue described in 5.1.1 and SpO_2 was not considered since its worst value is always equal to the minimum.

B.1.8 Statistically significant differences

These are features that were also adapted from Chen et al. [118], and they consist in assessing if differences between two windows are statistically significant. In this thesis, the two windows compared were the one corresponding to the initial 3 hours of data and the one corresponding to the most recent 3 hours of data, both belonging to the 12-hours window being analyzed.

The four tests implemented to assess those differences were: Wilcoxon rank sum test, two-sample Kolmogorov-Smirnov test, two-sample t-test and the two-sample F-test. The test statistics, and respective p-values, obtained from the four tests were used as features. These were extracted from the four vital signs and the QRSa time series.

B.1.9 Quadratic and cubic fits

Accordingly to Chen et al. [118], a quadratic regression was fitted to the data and four of its parameters were kept as features: first order coefficient, second order coefficient, R-squared and residuals variance. This was applied to the four vital signs and to the QRSa time series.

In addition to that, a cubic regression was fitted to the most recent 20-minutes window of data, accordingly to Lara et al. [154], and four of its parameters were kept as features: zero order coefficient, first order coefficient, second order coefficient and third order coefficient. This was only applied to the four vital signs.

B.1.10 Vital signs transient features

The first transient feature, *transient_trend*, is meant to be an indication of the vital signs behavior in the most recent 15-minutes window. This is, it denotes if the respective vital sign is increasing, decreasing or constant, with respect to a predefined threshold:

$$transient\_trend = \begin{cases} 1 \ \ (increasing), & m \geq r \\ 0 \ \ (constant), & |m| < |r| \\ -1 \ \ (decreasing), & m \leq -r \end{cases} \tag{B.3}$$

where m is the 15-minutes window slope, which had been calculated before (see B.1.4), and r is a threshold set to $\tan(15°)$, accordingly to Lara et al. [154].

The second transient feature, *magnitude_change*, complements the first one, since it indicates the vital sign's magnitude of change in the window being considered. This is accomplished by estimating the maximum deviation between the beginning and the end of the window, as given by:

$$magnitude\_change = max\left\{|max(S_p^+) - min(S_p^-)|, |max(S_p^-) - min(S_p^+)|\right\} \tag{B.4}$$

where the window being considered corresponds to the most recent 20 minutes, S_p^- is a window's subset which contains all samples between t_{min} and $t_{min} + (t_{max} - t_{min})p$ and S_p^+ is a window's subset which contains all samples between $t_{min} + (t_{max} - t_{min})(1-p)$ and t_{max}. t_{min} is the first window's timestamp and t_{max} is the last window's timestamp. p is a value between 0 and 1 and represents a percentage of the window's duration. $p = 0.2$ was selected, accordingly to Lara et al. [154].

Both these features were adapted from Lara et al. [154], where a more detailed explanation regarding these features can be found.

B.1.11 Maximum average temperature

In the reliability study performed by Stuiver et al. [126], regarding the temperature sensor used, it was suggested that hourly measures of BTemp provide the fundamental information needed to identify deterioration in terms of BTemp changes.

Therefore, and to simulate this acquisition process, all measurements in the same hour were averaged. This transforms the continuous 12-hours time series of BTemp data into 12 samples. Then, the maximum amongst them was stored as feature. The minimum was not considered due to the sensor issue described in 5.1.1.

B.1.12 Vital signs ticks

Upticks and downticks intend to reflect large changes in some variable's values between samples acquired in a short period of time. These definitions were adapted from Chen et al. [118], and, for this thesis, a large change was considered an absolute change ≥ 5% of the variable's physiological range, between measurements separated by 2 minutes or less.

The implemented physiological ranges were obtained by subtracting the lower thresholds to the upper thresholds, described in table 5.1. Given the range, the number of upticks and downticks for each vital sign can be determined by:

$$Uptick(dif) = \begin{cases} 0, & dif/r < 0.05 \\ 1, & dif/r \geq 0.05 \end{cases} \qquad NumUpT = \sum_i Uptick(dif_i) \tag{B.5}$$

$$Downtick(dif) = \begin{cases} 0, & dif/r > -0.05 \\ 1, & dif/r \leq -0.05 \end{cases} \qquad NumDownT = \sum_i Downtick(dif_i) \tag{B.6}$$

where r is the vital sign physiological range, dif is the difference between two samples and the summation on i regards every pair of samples separated by 2 minutes or less in the vital sign time series.

Besides $NumUpT$ and $NumDownT$, a third feature was extracted: $\frac{NumUpT}{NumUpT+NumDownT}$. These were not extracted from the BTemp time series, due to the sensor issue described in 5.1.1.

B.1.13 Correlations between vital signs

Correlations between vital signs (including auto-correlation) were calculated using the most recent 1-hour windows of data from the four vital signs, as performed by Lehman et al. [124].

B.1.14 Top-10 wavelet coefficients

For each vital sign, the most recent 1-hour window of data was used to extract the top-10 coefficients of the discrete wavelet transform using the Daubechies-4 wavelet, as performed by Sun et al. [127]. These coefficients were kept as features.

B.1.15 Histogram of derivatives

This procedure involves determining the distribution of the first and second order derivatives of the HR and RR signals. The goal is to get the frequency distribution of change in signal's intensity.

To achieve that, the first and second derivatives of each signal were computed. Then, 20-bin frequency histograms were constructed. The histogram limits were determined by analyzing the entire dataset, after outliers removal. This is, the maximum and minimum

values found for each derivative, across the entire dataset, were defined as the histogram limits.

This feature extraction procedure results in 40 coefficients per signal (20 from the first derivative histogram and 20 from the second derivative histogram).

This process was highly inspired by Dal Canton's work [106], where a more detailed explanation on this feature extraction method can be found.

B.1.16 RRI time domain features

Several RRI time domain features were extracted. Some of them were already described in table B.1, while the remaining are summarized in table B.2. It's worthy to reiterate that these features were all extracted from the RRI time series with trend.

Table B.2: Summary of the RRI time domain features extracted. Only the ones that had not been described yet (in table B.1) are displayed.

Feature	Description
sdann	Standard deviation of the average interbeat interval for each 5-minutes segment of the 12-hours window [66], [68]
sdnni	Average of the standard deviation of the interbeat intervals for each 5-minutes segment of the 12-hours window [66], [68]
*nn*50	Number of successive interbeat intervals that differ by more than 50 ms [67], [68]
*pnn*50	Percentage of successive interbeat intervals that differ by more than 50 ms [67], [68]
rmssd	Root mean squared sum of differences between successive interbeat intervals [66]–[68]
Number of outliers	Number of successive interbeat intervals that differ by more than 20% of the previous interbeat interval [67]
Percentage of outliers	Percentage of successive interbeat intervals that differ by more than 20% of the previous interbeat interval
hti	Heart rate variability triangular index
tinn	Triangular interpolation of the interbeat interval histogram

Both the heart rate variability triangular index, *hti*, and the triangular interpolation of the interbeat interval histogram, *tinn*, are based on the construction of the RRI time series histogram. A hypothetical RRI time series histogram is displayed in figure B.1, where $D(t)$ represents the RRI density distribution. Its maximum value, Y, located at

$t = X$, is used for the calculation of hti, as given by $hti = \frac{N_{RRI}}{Y}$, where N_{RRI} is the number of samples in the time series, which can be seen as the histogram area [68].

For the calculation of $tinn$, N and M (see figure B.1) have to be defined on the time axis. This is done by establishing a triangular function, $q(t)$, which satisfies $q(t) = 0$ for $t \leq N$ and $t \geq M$, and is obtained by minimizing $\int_0^{+\infty}(D(t) - q(t))^2 dt$. Then, $tinn$ is simply defined as $tinn = M - N$ [68].

Both these two features were extracted using a MATLAB toolbox available online [155].

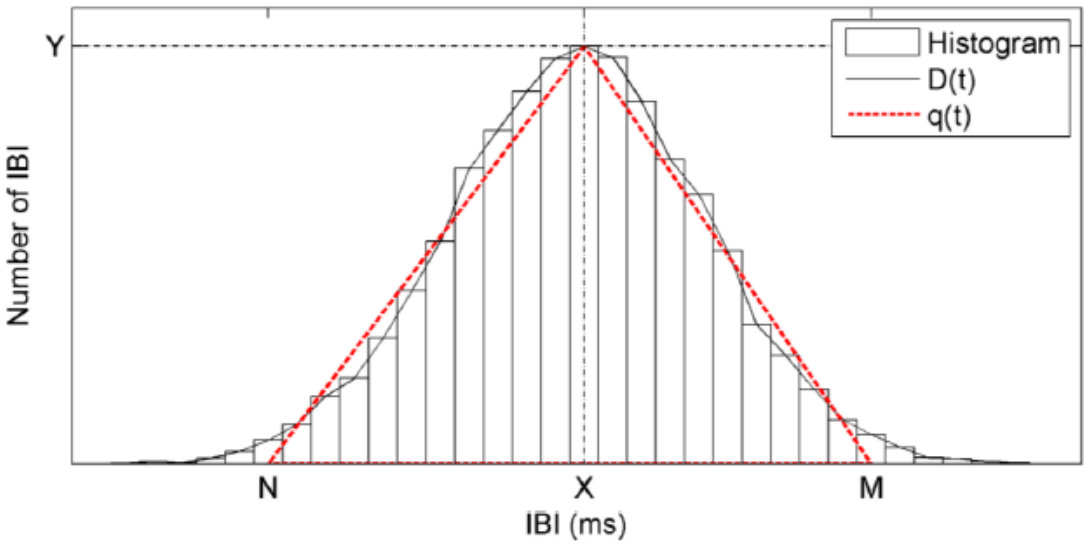


Figure B.1: Histogram of a hypothetical RRI time series. $D(t)$ is the samples' density distribution and $q(t)$ represents a triangular function fitted to $D(t)$ by minimizing the integral of the squared difference between $D(t)$ and $q(t)$. N and M represent the triangle base limits. $Y = D(X) = max(D(t))$. IBI - interbeat interval, which is equivalent to RRI. Withdrawn from [68].

B.1.17 RRI frequency domain features

Several frequency domain features were extracted from the RRI time series power spectrum. Since this was an unevenly sampled time series, and for the reasons presented in 5.3, the Lomb-Scargle periodogram was utilized. For a theoretical explanation of this technique, consult Clifford et al. [137].

Usually, for these features extraction, four frequency bands are distinguished in the power spectrum: ultra-low frequency ([0, 0.0033] Hz), very low frequency (]0.0033, 0.04] Hz), low frequency (]0.04, 0.15] Hz) and high frequency (]0.15, 0.4] Hz) [67], [68]. However, the ultra-low frequency band requires a recording of at least 24 hours [66]. Given that 12-hours windows were being used, this band was not considered.

Table B.3 provides a summary of the frequency domain features extracted from the other three bands. It's worthy to reiterate that these were all extracted considering the

detrended RRI time series.

Table B.3: Summary of the RRI frequency domain features extracted. Adapted from [67].

Feature	Description
vlf	Absolute power of the VLF band
lf	Absolute power of the LF band
hf	Absolute power of the HF band
$total$	Total absolute power of the VLF, LF and HF bands
$pvlf$	$pvlf = \frac{vlf}{total}$
plf	$plf = \frac{lf}{total}$
phf	$phf = \frac{hf}{total}$
plf_nu	$plf\_nu = \frac{lf}{total-vlf}$
phf_nu	$phf\_nu = \frac{hf}{total-vlf}$
$lfhf\_ratio$	$lfhf\_ratio = \frac{lf}{hf}$
$peakfreq\_vlf$	Frequency that presented the maximum power in the VLF band
$peakfreq\_lf$	Frequency that presented the maximum power in the LF band
$peakfreq\_hf$	Frequency that presented the maximum power in the HF band
Abbreviations: VLF - very low frequency, LF - low frequency, HF - high frequency	

B.1.18 RRI non-linear features

The RRI non-linear features extracted were of three types: based on entropy measures, based on the Poincaré plot analysis and based on fractal measures. It's worthy to reiterate that all of them were extracted from the RRI time series with trend.

Entropy-based measures

The first entropy-based measure extracted was approximate entropy, *apen*. This is a measure of regularity and complexity of a time series. It can also be interpreted as the probability that similar patterns won't repeat in the time series [156]. Large *apen* values

indicate a RRI time series high in entropy and with low predictability, while small values correspond to a predictable and regular signal [66].

For this measure's calculation, two parameters have to be determined: the dimension parameter, m, and the filter parameter, r. Accordingly to Batchinsky et al. [69], $m = 2$ and $r = 0.2\sigma$, were selected. σ represents the time series' standard deviation. A more detailed explanation on this feature's calculation can be found in Pincus' work [157].

In practice, a MATLAB toolbox available online was used to extract this feature [155]. Given the results obtained by Batchinsky et al. [69], this was performed on the most recent 1000-samples window.

The second entropy-based measure extracted was sample entropy, *spen*. This measure's interpretation is very similar to the one made for *apen*, however *spen* provides a less biased and more reliable estimation of the time series' complexity [66].

It also requires the same two parameters to be determined, which were attributed the same values as for *apen*, accordingly to Batchinsky et al. [69]. The most recent 1000-samples window was also used here. A more detailed explanation on this feature's calculation can be found elsewhere [68], [158].

In practice, a MATLAB toolbox available online was also used to extract this feature [159].

Poincaré plot analysis

A Poincaré plot, also known as return map, is a plot of every RRI sample against the respective previous sample in the time series. Its analysis quantifies self-similarity [68] and allows the visual identification of hidden patterns in the time series and the extraction of several measures [66]. For the latter, an ellipse is usually fitted to the plotted data, as in figures B.2. Then, four non-linear measures can be calculated (1) S, which is the ellipse area and it represents total heart rate variability [66]; (2) $SD1$, which is the standard deviation along the line perpendicular to the line of identity and passing through the mean of the RRI time series (see figure B.2). It corresponds to the ellipse's width and represents short term variability [68]; (3) $SD2$, which is the standard deviation along the line of identity (see figure B.2). It corresponds to the ellipse's length and represents long term variability [68]; (4) $SD1/SD2$, which is the ratio between $SD1$ and $SD2$, and represents the RRI time series unpredictability [66].

In this thesis, the above-mentioned features were extracted considering the entire 12-hours window and using a MATLAB toolbox available online [155].

Figure B.2 (a) shows the Poincaré plot of a 0-labeled window, while figure B.2 (b) shows the Poincaré plot of a 1-labeled window. Both figures are in line with what would be expected, since healthy subjects typically present this dispersed "comet"-like shape (B.2 (a)) and deteriorating patients might present atypical shapes, like the concentrated "torpedo"-like shape observed in figure B.2 (b) [160]. In addition to that, both $SD1$ and $SD2$ are reduced for the deteriorating-case window, which indicates a decrease in the RRI

time series complexity. This was also expected, since a reduction in this complexity is associated with complications development, as discussed in subsection 2.3.2.

Fractal-based measures

Fractal measures are based on the concept that a system can be fractioned into smaller parts, where each of them resembles one another but on a different scale. Detrended fluctuation analysis intends to quantify the fractal properties of a non-stationary time series, which means the changing scale here is time. To achieve that, fluctuations of an integrated and detrended time series are measured at different scales and plotted against the scale's size [68]. By considering two different regions on this plot, two features can be calculated: α_1, which represents short-term fluctuations, and α_2, which represents long-term fluctuations [66]. In this thesis, these two regions were separated at the 16 samples scale. A more detailed explanation on these features calculation can be found in Rashmur's work [68].

In practice, the entire 12-hours window was considered and a MATLAB toolbox available online was used to extract these features [155].

B.1.19 RRI non-linear vector map features

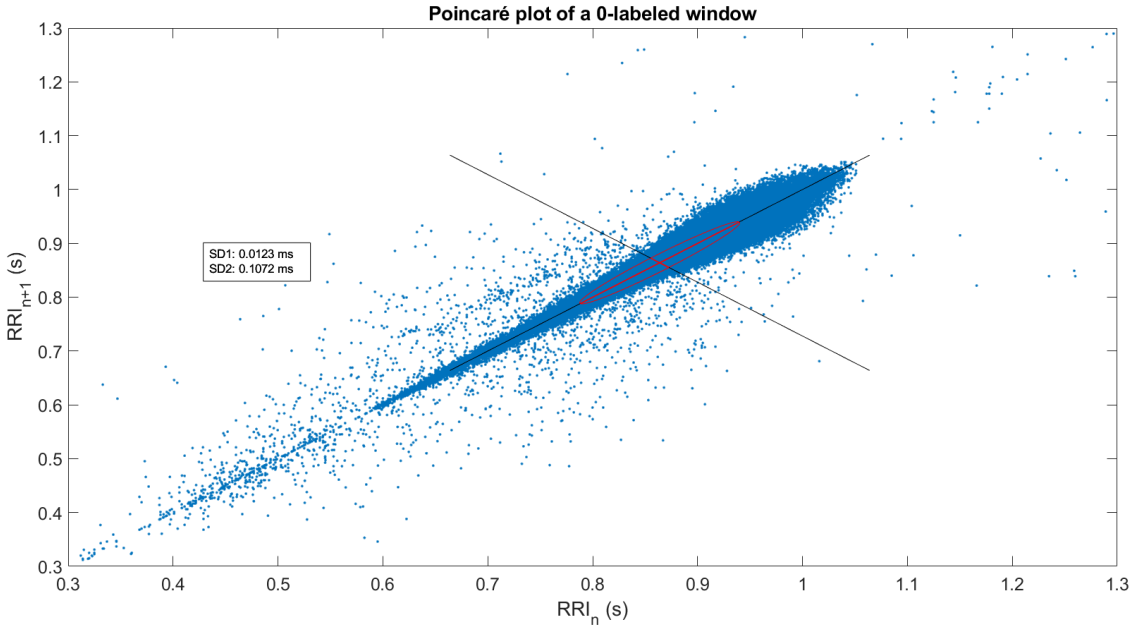
Besides the already described features extracted from the Poincaré plot, Moridani et al. [161] developed three more. These were based on the idea that subjects with similar $SD1$ and $SD2$ might present different temporal dynamics in the Poincaré plot. This is, they connected every two consecutive points with a vector and developed features to analyze those vectors. By performing this assessment, the temporal dynamics of the Poincaré plot construction is better described.

In practice, the three features were extracted in two different ways. The first was implemented due to a result obtained by Moridani et al. [161], and it consists in considering only the most recent 30 minutes segment, to extract the mean and the standard deviation of the feature. The second calculates the mean considering 30 minutes segments at a time, traversing the entire 12-hours window. Both these extraction methods are slightly different from the one reported by Moridani et al. [161].

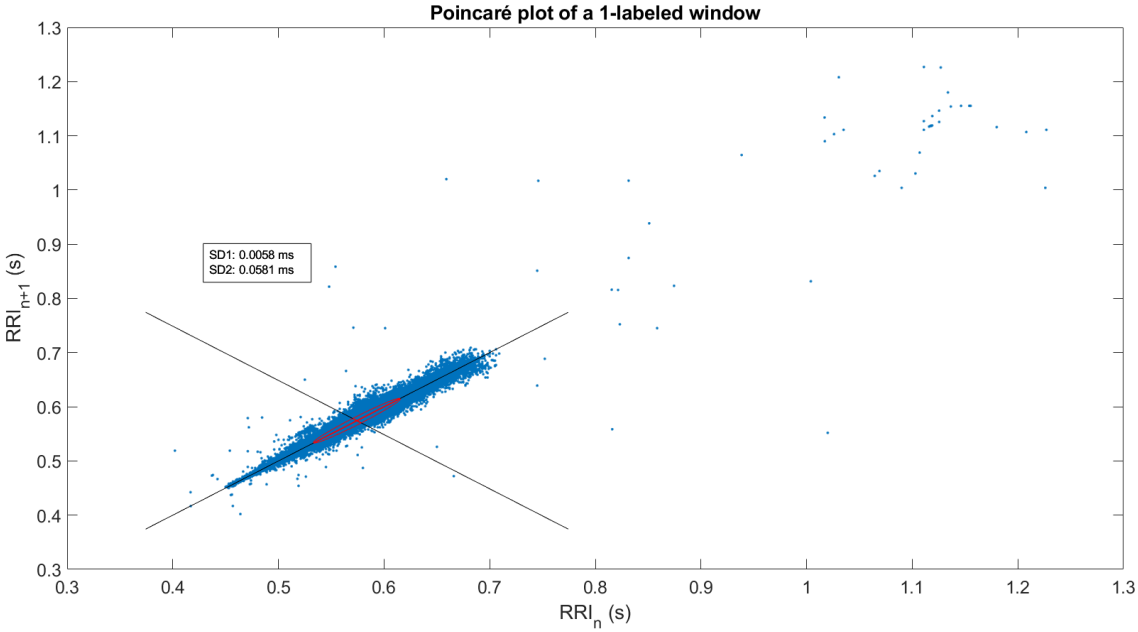
For a more detailed explanation than the one provided next, consult Moridani et al. [161].

Angle between vectors

In this feature, the angle between consecutive vectors is calculated. This has some limitations, since vectors with very distinct magnitudes can present the same angle in-between them.



(a) Poincaré plot of a 0-labeled window.



(b) Poincaré plot of a 1-labeled window.

Figure B.2: Poincaré plots of two windows with different labels. The ellipse's center is located at $(mean(RRI), mean(RRI))$, where $mean(RRI)$ represents the mean of the RRI time series being considered. The ellipse's width and length correspond to $SD1$ and $SD2$, respectively. The two black line segments correspond to the identity line ($y = x$) and to the line perpendicular to the identity line that crosses the ellipse's center, $(mean(RRI), mean(RRI))$. RRI_n is the RRI at time n, whereas RRI_{n+1} is the RRI at time $n+1$.

Area of the triangle made by successive points

To solve the limitation of the previous feature, the area of the triangle made by successive points was introduced. This consists in calculating the area of the triangle formed by the two vectors and the connection between the origin of the first vector and the end of the second.

Shortest distance to the identity line

The shortest distance from each point to the identity line can provide valuable information, since it exposes that heart rate variability might be decreasing. This distance corresponds to the length of a straight line drawn between the point being considered and the identity line.

B.2 Ordinal/categorical features

B.2.1 Age coefficient

The age coefficient, age_coef, intends to assign a more severe score for older subjects and corresponds to:

$$age\_coef = \begin{cases} 0, & age < 40 \\ 1, & 40 \leq age \leq 75 \\ 2, & age > 75 \end{cases} \tag{B.7}$$

where age is the subject's age. The thresholds selection was guided by previous studies [103], [119].

B.2.2 Vital signs categorical coefficients

The vital signs categorical coefficients were calculated considering only the most recent 3 hours of data belonging to the window. These are EWS-like scores but, unlike EWS, these coefficients were personalized by considering the subject's age when attributing the scores. For HR, BTemp and SpO_2, age only contributed to assign a more or less severe score. However, for RR, age contributed to readjust the thresholds.

The vital signs categorical coefficients assignment is detailed in table B.4. The thresholds selection was guided by previous studies [103], [114], [119], [162].

B.2.3 Partial dependencies

Partial dependencies are three EWS-like coefficients that explore relations between HR, RR and age. These were implemented by analyzing partial dependency plots reported by Churpek et al. [78], where interactions between the risk of developing an AE and pairs of these variables were presented.

Table B.4: Vital signs categorical coefficients calculation. $HR\_catcoef$, $RR\_catcoef$, $Temp\_catcoef$ and $SpO2\_catcoef$ are the categorical coefficients for HR, RR, BTemp and SpO_2, respectively. hr, rr, $temp$ and $spo2$ are the mean in the most recent 3 hours for HR, RR, BTemp and SpO_2, respectively, in the same units as previously reported. age is the subject's age.

Coefficient	if $age < 65$	if $65 \leq age < 80$	if $age \geq 80$
$HR\_catcoef$	$\begin{cases} 2, & hr < 40 \\ 0, & 40 \leq hr \leq 100 \\ 1, & 100 < hr \leq 140 \\ 2, & hr > 140 \end{cases}$	$\begin{cases} 3, & hr < 40 \\ 0, & 40 \leq hr \leq 100 \\ 2, & 100 < hr \leq 140 \\ 3, & hr > 140 \end{cases}$	$\begin{cases} 4, & hr < 40 \\ 0, & 40 \leq hr \leq 100 \\ 3, & 100 < hr \leq 140 \\ 4, & hr > 140 \end{cases}$
$RR\_catcoef$	$\begin{cases} 2, & rr < 8 \\ 1, & 8 \leq rr < 16 \\ 0, & 16 \leq rr \leq 20 \\ 1, & 20 < rr \leq 30 \\ 2, & rr > 30 \end{cases}$	$\begin{cases} 2, & rr < 12 \\ 0, & 12 \leq rr \leq 28 \\ 2, & rr > 28 \end{cases}$	$\begin{cases} 2, & rr < 10 \\ 0, & 10 \leq rr \leq 30 \\ 2, & rr > 30 \end{cases}$
$Temp\_catcoef$	$\begin{cases} 1, & temp < 35 \\ 0, & 35 \leq temp \leq 38 \\ 1, & temp > 38 \end{cases}$	$\begin{cases} 2, & temp < 35 \\ 0, & 35 \leq temp \leq 38 \\ 2, & temp > 38 \end{cases}$	$\begin{cases} 3, & temp < 35 \\ 0, & 35 \leq temp \leq 38 \\ 3, & temp > 38 \end{cases}$
$SpO2\_catcoef$	$\begin{cases} 5, & spo2 < 78 \\ 4, & 78 \leq spo2 < 82 \\ 3, & 82 \leq spo2 < 86 \\ 2, & 86 \leq spo2 < 90 \\ 1, & 90 \leq spo2 < 95 \\ 0, & spo2 \geq 95 \end{cases}$	$\begin{cases} 6, & spo2 < 78 \\ 5, & 78 \leq spo2 < 82 \\ 4, & 82 \leq spo2 < 86 \\ 3, & 86 \leq spo2 < 90 \\ 1, & 90 \leq spo2 < 95 \\ 0, & spo2 \geq 95 \end{cases}$	$\begin{cases} 7, & spo2 < 78 \\ 6, & 78 \leq spo2 < 82 \\ 5, & 82 \leq spo2 < 86 \\ 4, & 86 \leq spo2 < 90 \\ 2, & 90 \leq spo2 < 95 \\ 0, & spo2 \geq 95 \end{cases}$

These coefficients were also obtained considering only the most recent 3 hours of data belonging to the window, and are calculated as:

$$pd6\_coef = \begin{cases} 0, & hr < 110 \wedge 18 \leq rr \leq 22 \\ 1, & hr < 110 \wedge rr < 18 \\ 2, & hr \geq 110 \wedge rr < 22 \\ 3, & rr > 22 \end{cases} \quad pd7\_coef = \begin{cases} 0, & hr \leq 100 \wedge age \leq 50 \\ 1, & hr \leq 100 \wedge age > 50 \\ 2, & hr > 100 \wedge age \leq 50 \\ 3, & hr > 100 \wedge age > 50 \end{cases} \quad pd9\_coef = \begin{cases} 0, & 18 \leq rr \leq 22 \wedge age \leq 50 \\ 1, & 18 \leq rr \leq 22 \wedge age > 50 \\ 2, & rr < 18 \wedge age \leq 50 \\ 3, & rr < 18 \wedge age > 50 \\ 4, & rr > 22 \end{cases} \tag{B.8}$$

where $pd6\_coef$, $pd7\_coef$ and $pd9\_coef$ are the partial dependencies coefficients. hr and rr are the mean in the most recent 3 hours for HR and RR, respectively, in the same units as previously reported. age is the subject's age.

B.2.4 ASA

This feature is simply the American Society of Anesthesiologists class (ASA). This is a classification system that categorizes patients accordingly to their preoperative physiological

status and it considers 6 different classes [135].

B.2.5 Number of comorbidities

11 types of comorbidities, plus the presence of others, were included in the collected dataset, as previously mentioned. Therefore, this feature, num_comorb, can take integer values and ranges from 0 to 12.

B.2.6 Multiple comorbidities

This feature is defined as 1, if $num\_comorb > 1$, and 0, otherwise.

B.2.7 RRI tree-based rules

These features were derived from tree-based rules obtained by Jovic et al. [156]. In their study, several features, extracted from 5 minutes RRI recordings, were combined with ML algorithms for the classification of ECG signals, based on their heart rate variability. Therefore, these rules address the presence of arrhythmias and other heart conditions.

The features calculation, accordingly to the obtained rules, is given by:

$$rri\_rule1 = \begin{cases} 0, & hti \leq 20.42 \wedge rmssd \leq 0.068 \\ 2, & hti > 20.42 \\ 1, & otherwise \end{cases} \qquad rri\_rule2 = \begin{cases} 0, & rrintstd > 0.038 \wedge rmssd \leq 0.056 \\ 2, & rrintstd \leq 0.038 \\ 1, & otherwise \end{cases} \tag{B.9}$$

where rri_rule1 and rri_rule2 are the two derived features. hti and $rmssd$ have the same meaning as before (see B.1.16), but, for this feature, were calculated in the most recent 5 minutes of RRI data. $rrintstd$ is the RRI standard deviation calculated in the most recent 5 minutes of data.

Appendix C

Results appendix

This appendix contains figures and tables that, despite relevant, were not considered for inclusion in the main text, mostly because of their dimensions.

Figures C.1 to C.6 refer to the assessment of correlations between the features considered for the clustering model development.

Figures C.7 to C.9 illustrate how linear interpolation, new approach version 1 and new approach version 2 would handle the same gap in a vital sign time series. The vital sign considered for the exemplification was respiration rate.

Figures C.10 to C.13 are the results of the error study performed considering only one vital sign at a time.

Figures C.14 and C.15 are the results of the error studies that sought to help define an appropriate past interval of QRSa data to apply the average and median techniques to. This was related with the handling of missing data periods in the QRSa time series.

Figures C.16, C.17 and C.18 represent a comparison example between a RRI time series preprocessed with the novel technique (described in 5.3.1.1) and the same time series not preprocessed, preprocessed with only the selective median filter and preprocessed with only the impulse rejection filter, respectively.

Figure C.19 is the result of assessing the differences in the computational time required by new approach version 1, new approach version 2 and linear interpolation, to handle missing data periods of different durations. 95% confidence intervals were calculated but were so small that are not visible in the figure. These simulations were conducted on a Lenovo Legion Y520 (CPU: Intel i5-7300HQ 2.50GHz, RAM: 8GB).

Table C.1 is a summary of features importance in the best model with initial features set 'NoTemp&SpO_2', based on the associated regression coefficients absolute values.

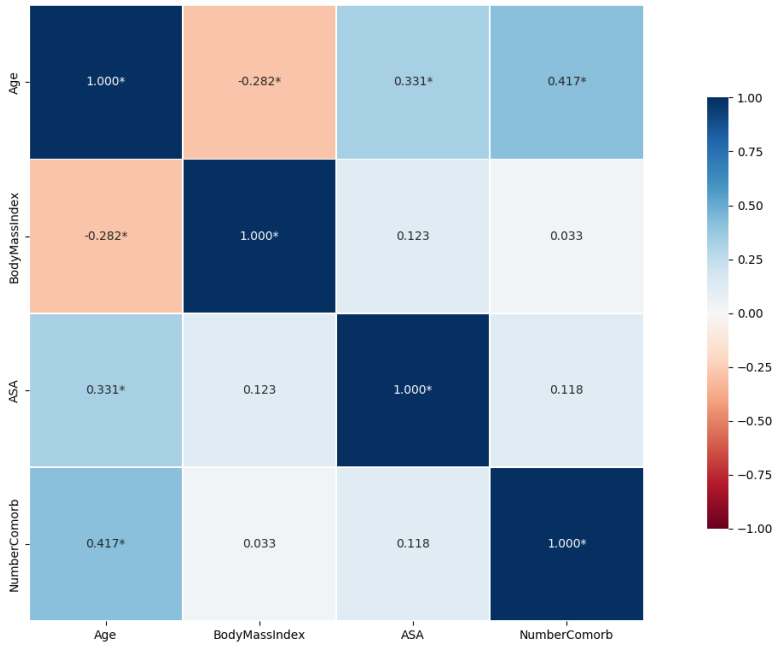


Figure C.1: Numerical-numerical features correlations assessment, for the clustering model development, using the Pearson correlation coefficient. *NumberComorb* is the number of comorbidities feature. * indicates statistically significant correlations, at 5% significance level.

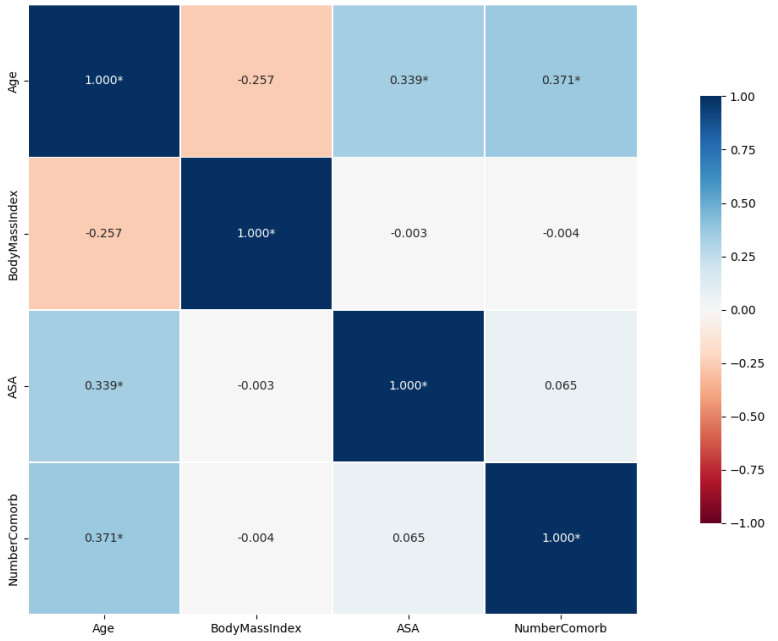


Figure C.2: Numerical-numerical features correlations assessment, for the clustering model development, using the Spearman rank correlation coefficient. *NumberComorb* is the number of comorbidities feature. * indicates statistically significant correlations, at 5% significance level.

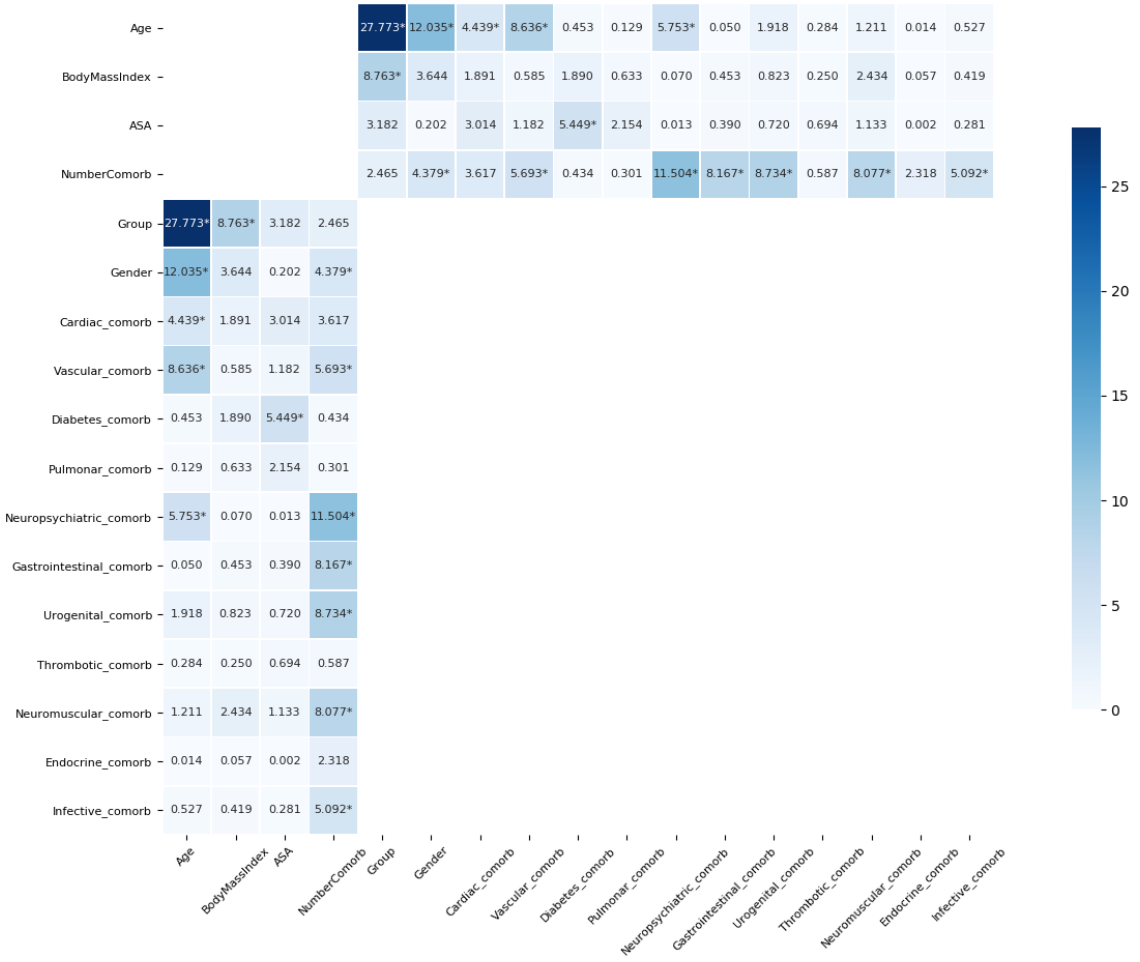


Figure C.3: Numerical-categorical features correlations assessment, for the clustering model development, using the Kruskal Wallis H test. *NumberComorb* is the number of comorbidities feature. Features that end with *_comorb* are features that refer to the presence or not of the respective type of comorbidity. * indicates statistically significant correlations, at 5% significance level.

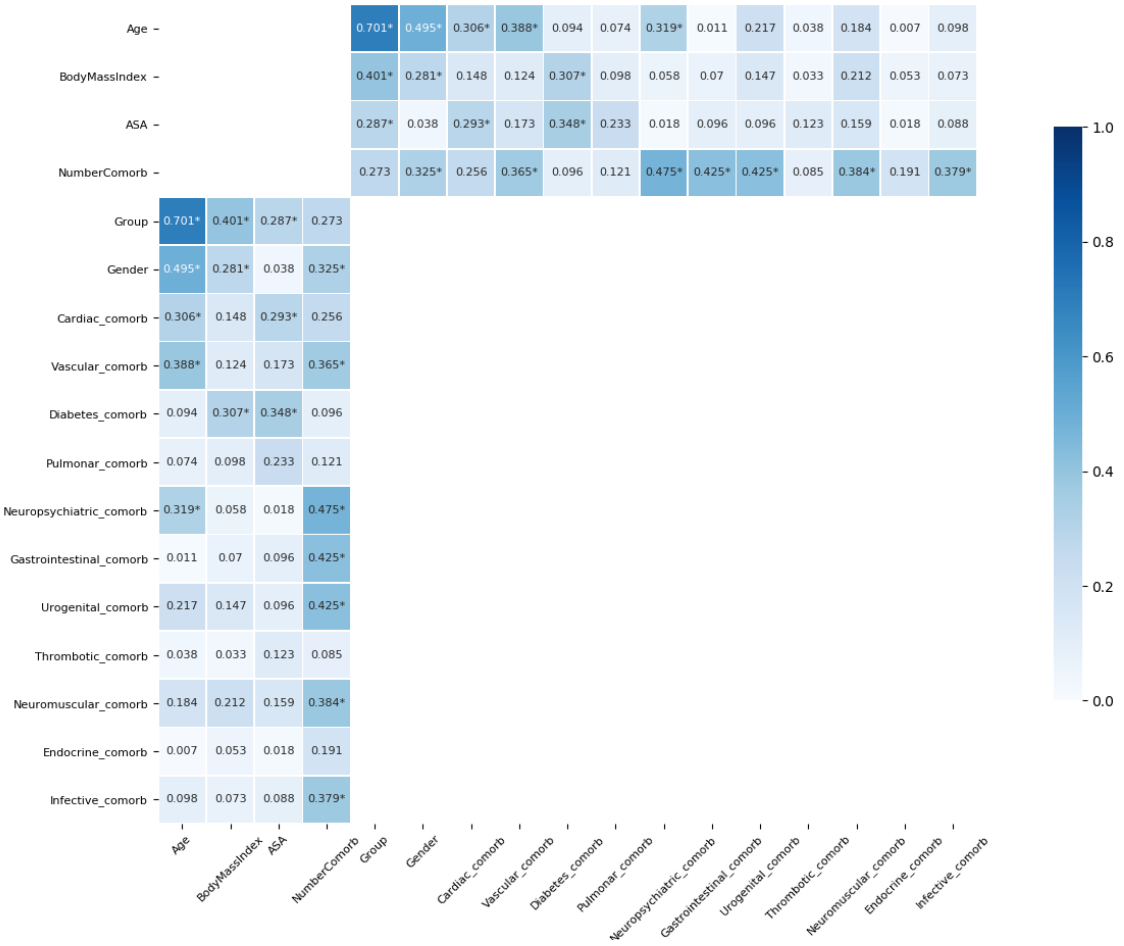


Figure C.4: Numerical-categorical features correlations assessment, for the clustering model development, using the eta correlation coefficient. *NumberComorb* is the number of comorbidities feature. Features that end with *_comorb* are features that refer to the presence or not of the respective type of comorbidity. * indicates statistically significant correlations, at 5% significance level.

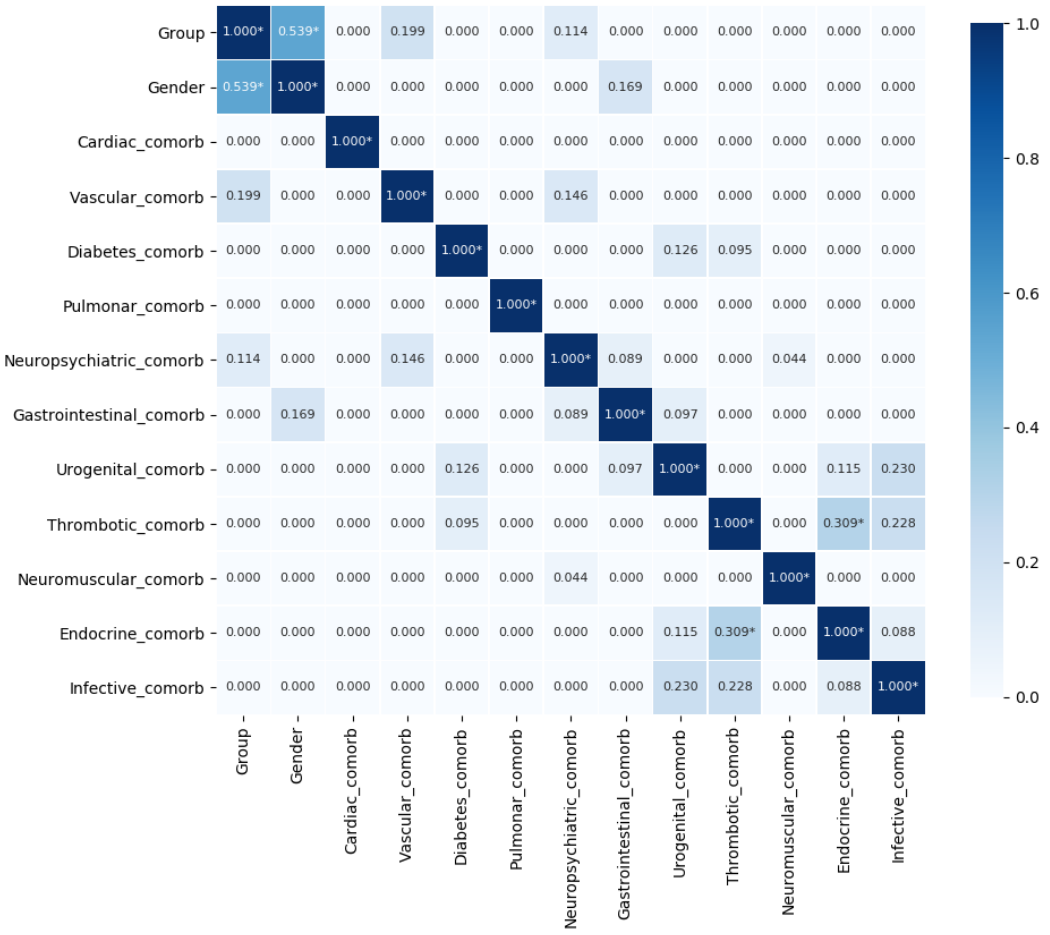


Figure C.5: Categorical-categorical features correlations assessment, for the clustering model development, using the Cramer's V coefficient. Features that end with *_comorb* are features that refer to the presence or not of the respective type of comorbidity. * indicates statistically significant correlations, at 5% significance level.

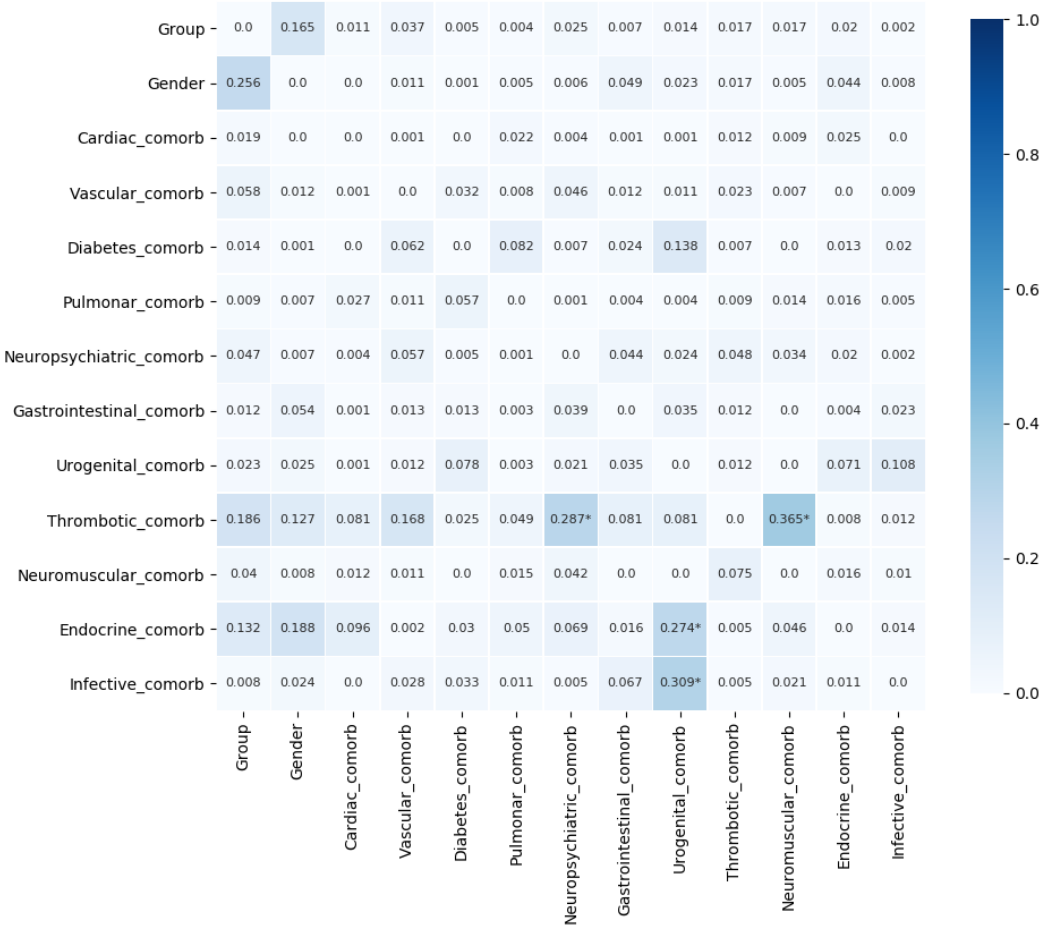


Figure C.6: Categorical-categorical features correlations assessment, for the clustering model development, using the Theil's U coefficient. Features that end with *_comorb* are features that refer to the presence or not of the respective type of comorbidity. * indicates statistically significant correlations, at 5% significance level.

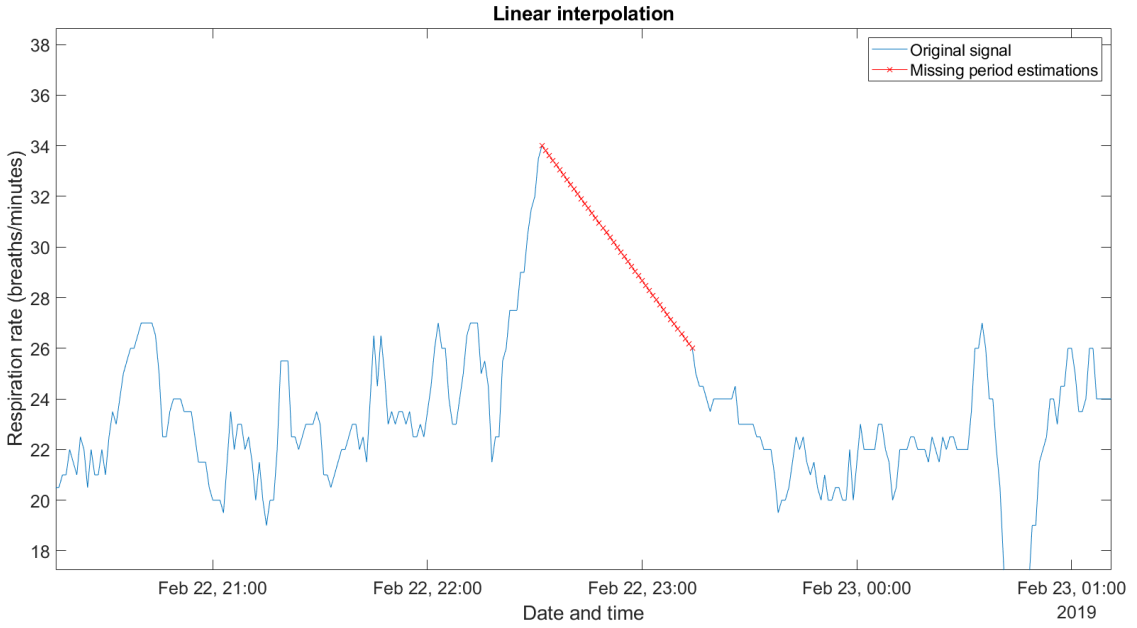


Figure C.7: Example of a missing data period, in a respiration rate time series, being handled using linear interpolation. The red crosses represent the estimated samples.

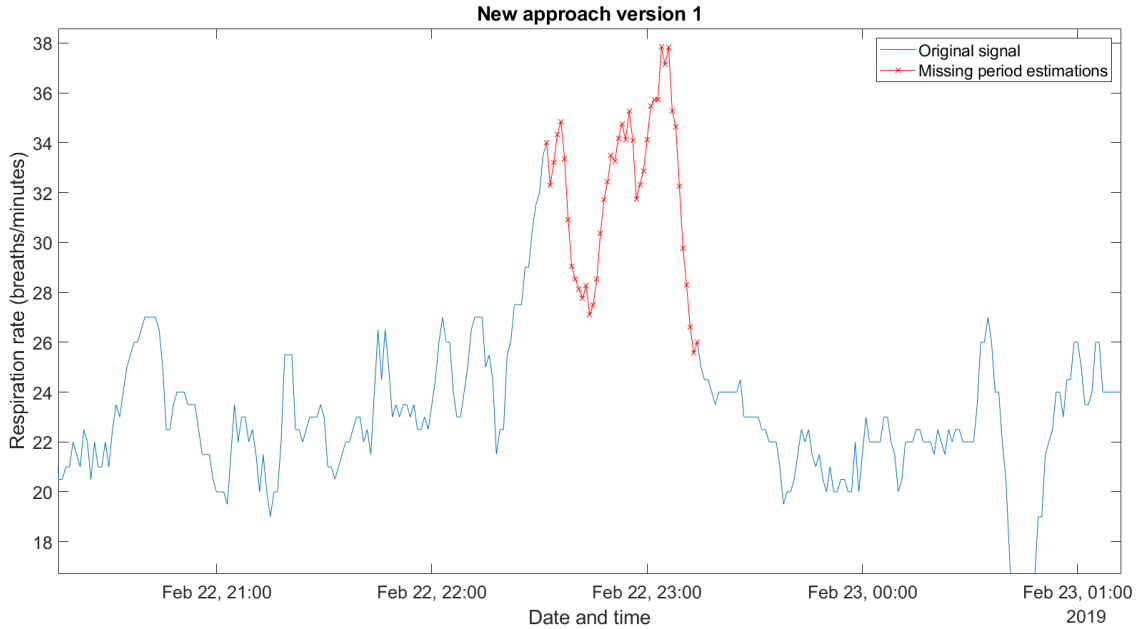


Figure C.8: Example of a missing data period, in a respiration rate time series, being handled using the new approach version 1. The red crosses represent the estimated samples.

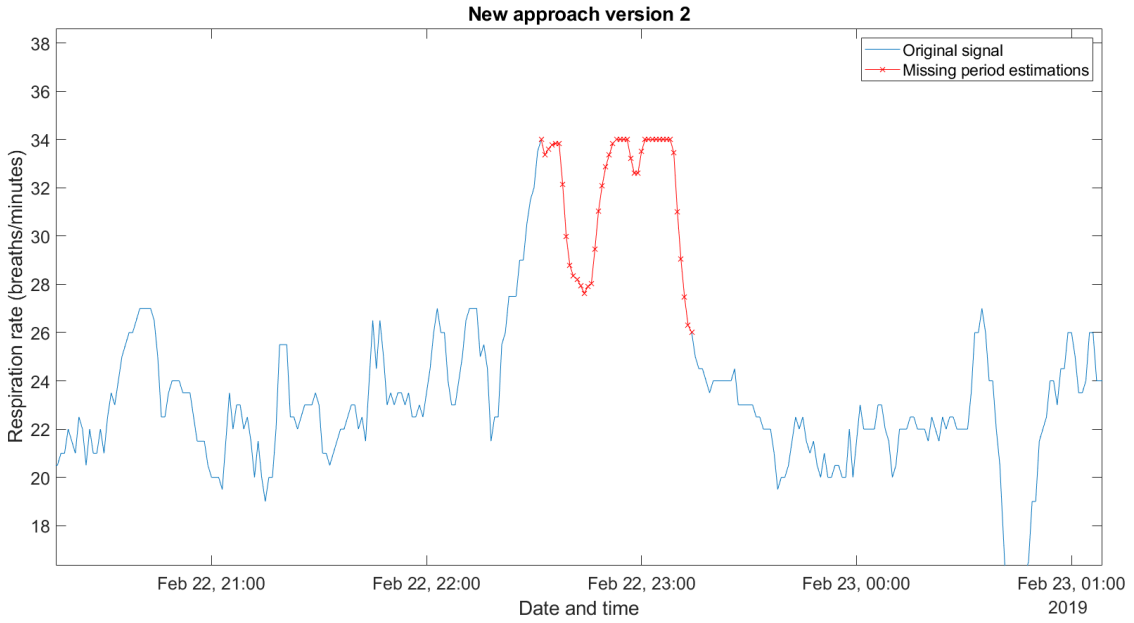


Figure C.9: Example of a missing data period, in a respiration rate time series, being handled using the new approach version 2. The red crosses represent the estimated samples.

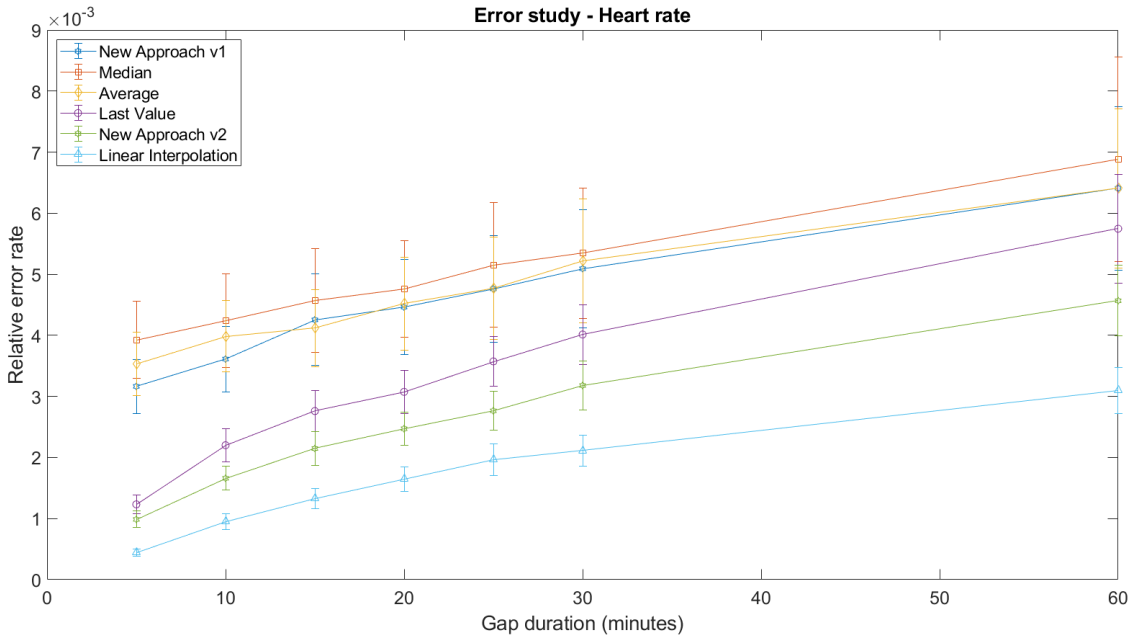


Figure C.10: Results of the error study performed for the selection of an adequate technique to handle missing data periods in the HR time series. This study's methodology and the six techniques being tested are described in 5.1.2.2. The error bars represent the 95% confidence interval.

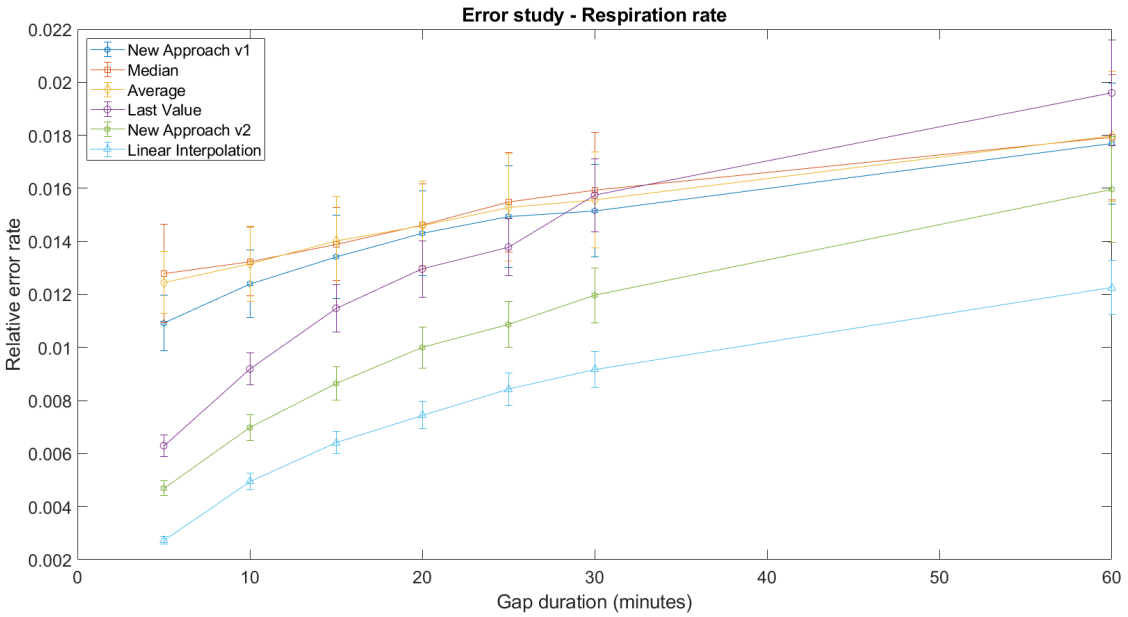


Figure C.11: Results of the error study performed for the selection of an adequate technique to handle missing data periods in the RR time series. This study's methodology and the six techniques being tested are described in 5.1.2.2. The error bars represent the 95% confidence interval.

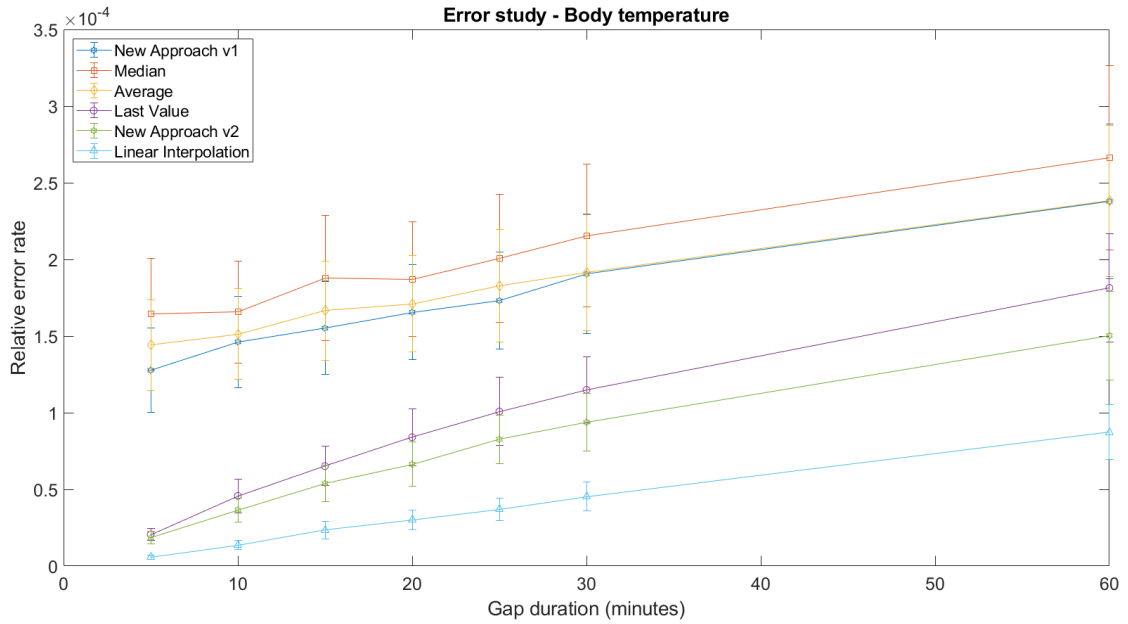


Figure C.12: Results of the error study performed for the selection of an adequate technique to handle missing data periods in the BTemp time series. This study's methodology and the six techniques being tested are described in 5.1.2.2. The error bars represent the 95% confidence interval.

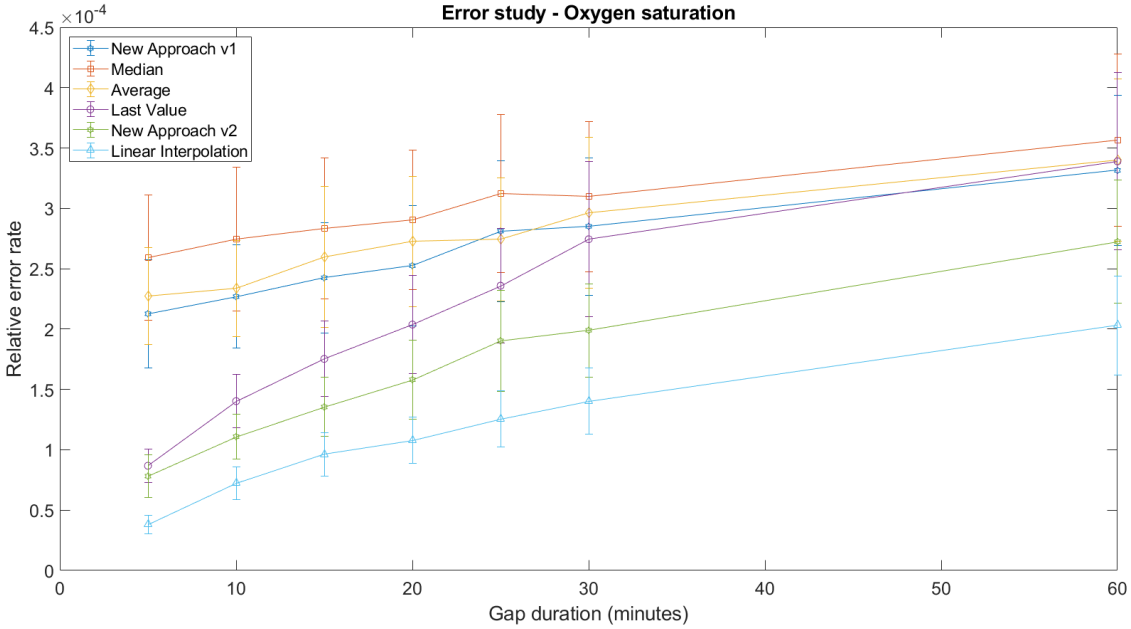


Figure C.13: Results of the error study performed for the selection of an adequate technique to handle missing data periods in the SpO_2 time series. This study's methodology and the six techniques being tested are described in 5.1.2.2. The error bars represent the 95% confidence interval.

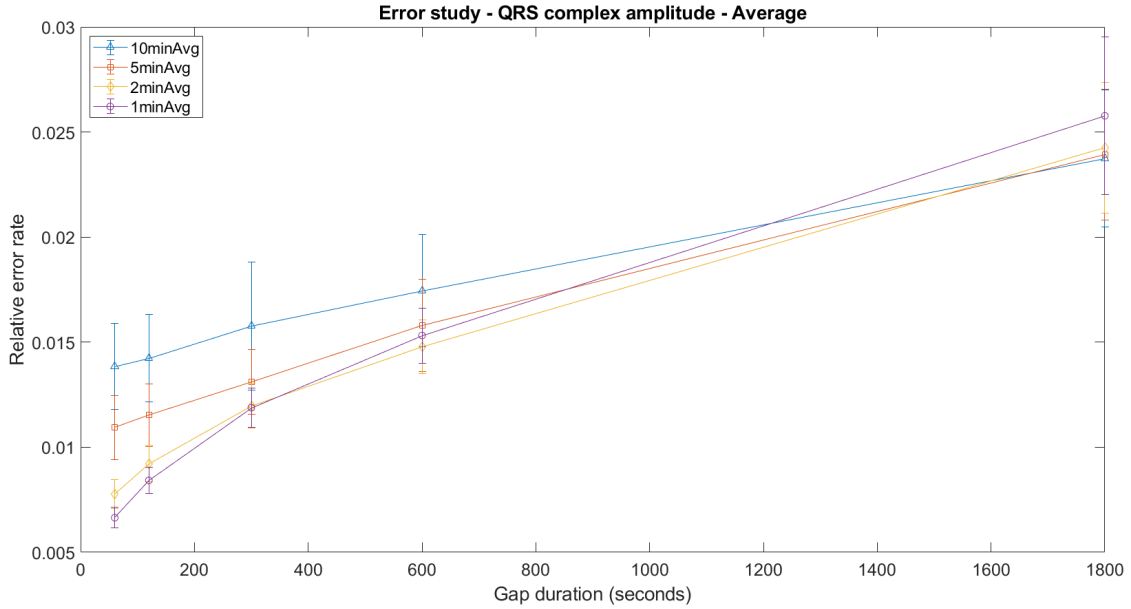


Figure C.14: Results of the error study performed for the selection of a suitable past interval to average over. This was executed for a proper implementation of the average technique, to test its handling of missing data periods in the QRSa time series. This study's methodology is described in 5.2.2.1. *NminAvg* refers to averaging over the past *N* minutes. The error bars represent the 95% confidence interval.

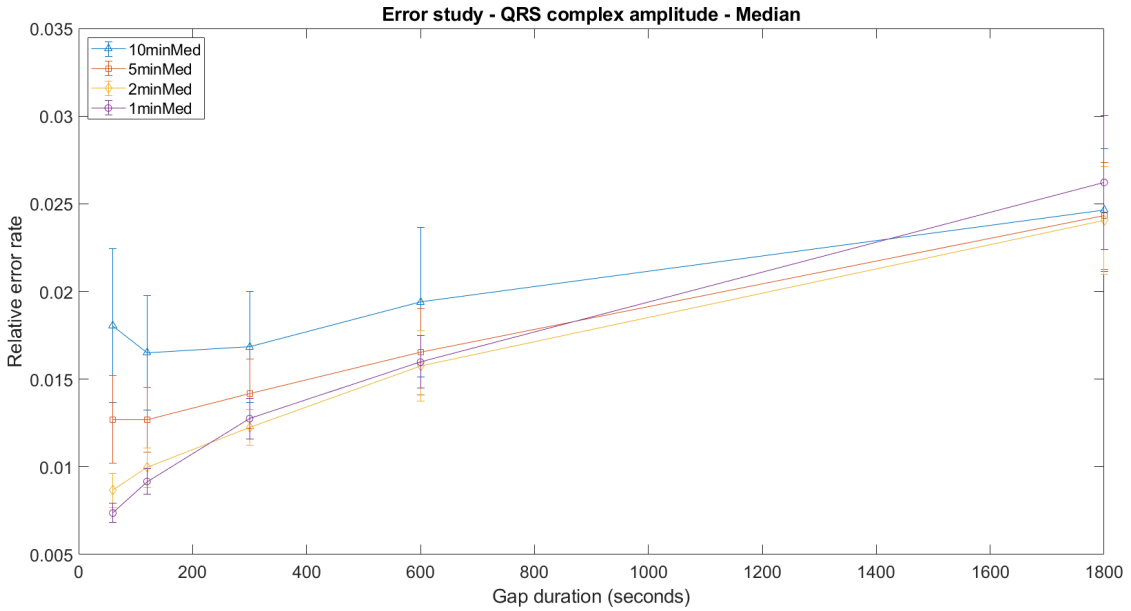


Figure C.15: Results of the error study performed for the selection of a suitable past interval to median over. This was executed for a proper implementation of the median technique, to test its handling of missing data periods in the QRSa time series. This study's methodology is described in 5.2.2.1. *NminMed* refers to applying the median over the past N minutes. The error bars represent the 95% confidence interval.

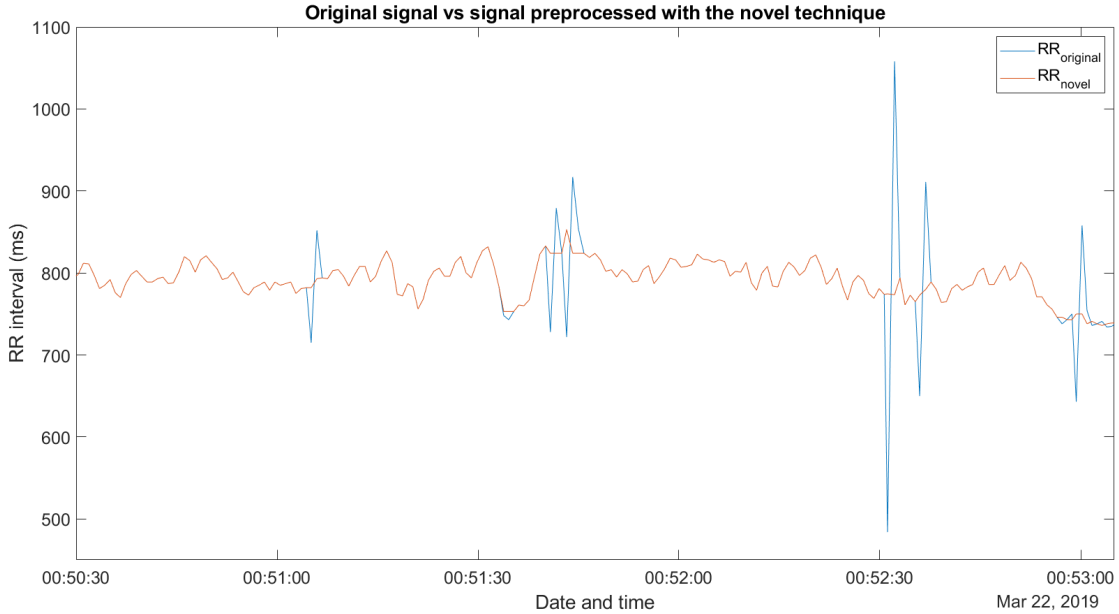


Figure C.16: Comparison example between an unprocessed RRI time series ($RR_{original}$) and the same time series preprocessed with the novel technique (RR_{novel}), which is described in 5.3.1.1. At least four ectopic beats can be identified in the unprocessed time series.

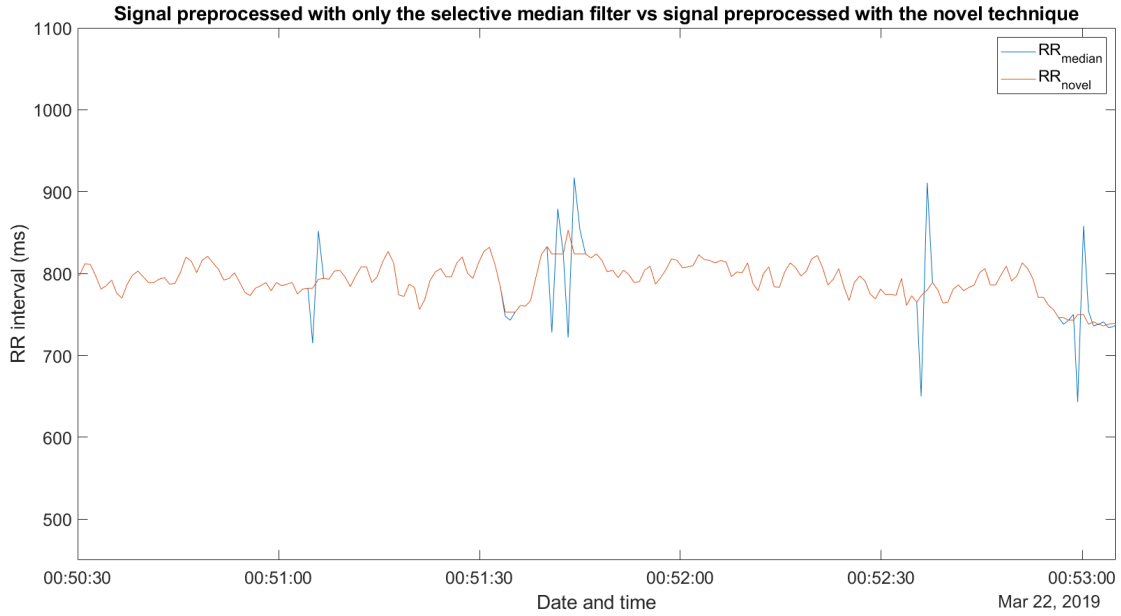


Figure C.17: Comparison example between a RRI time series preprocessed with only the selective median filter (RR_{median}) and the same time series preprocessed with the novel technique (RR_{novel}), which is described in 5.3.1.1.

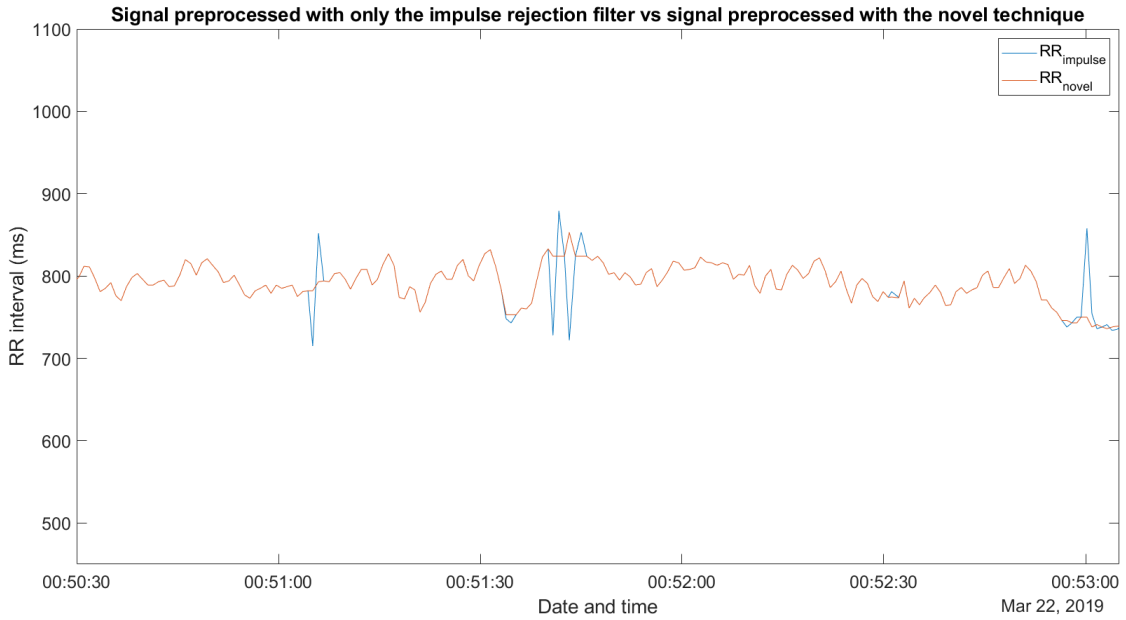


Figure C.18: Comparison example between a RRI time series preprocessed with only the impulse rejection filter ($RR_{impulse}$) and the same time series preprocessed with the novel technique (RR_{novel}), which is described in 5.3.1.1.

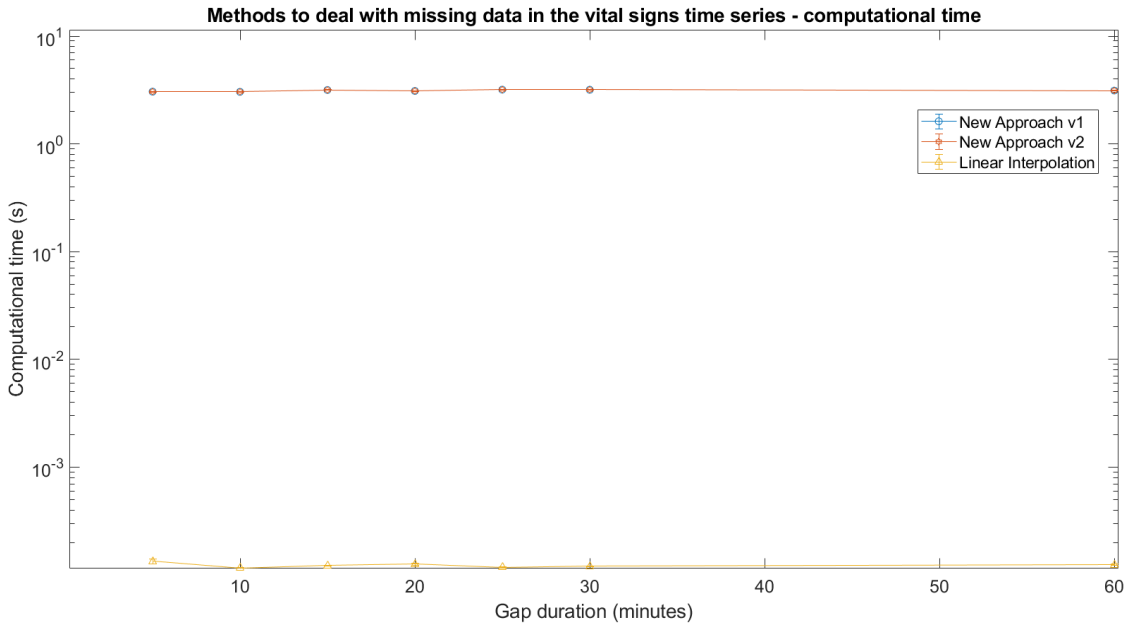


Figure C.19: Difference in the computational time required by new approach version 1, new approach version 2 and linear interpolation, to handle missing data periods of different durations. The techniques being tested are described in 5.1.2.2. Note that the computational time is presented in log scale.

Table C.1: Features importance in the best model with initial features set 'NoTemp&SpO_2', based on the associated regression coefficients absolute values. Features that are correlated with the respective deemed important feature are also presented.

Feature (explanation section)	Absolute regression coefficient value	Odds ratio[a]	Correlations
HR_catcoef_0 (B.2.2)	0.4079	0.6650	—
RR_catcoef_0 (B.2.2)	0.3718	0.6895	—
pd7_coef_1 (B.2.3)	0.3432	0.7095	—
corr_HR_RR (B.1.13)	0.2646	1.3029	RR wavelet coefficients, corr_RR_RR
HasMultipleComorbs_1 (B.2.6)	0.2632	0.7686	—
HRmed (B.1.7)	0.2275	1.2555	HRavg, RRIavg, HR_recentmin, RRImed, HR_exp_smooth_avg
HR_cubicfit_zerocoef (B.1.9)	0.2132	1.2376	HR wavelet coefficients, corr_HR_HR
age_coef_2 (B.2.1)	0.2005	0.8183	—
RR_normdiff (B.1.2)	0.1857	0.8305	—
HR_quadfit_firstcoef (B.1.9)	0.1822	1.1999	HR_quadfit_secondcoef
Hist_HR_32 (B.1.15)	0.1732	1.1891	—
age_coef_1 (B.2.1)	0.1688	0.8447	—
Hist_HR_12 (B.1.15)	0.1657	1.1802	—
RR_quadfit_rsquared (B.1.9)	0.1615	1.1753	—
pd9_coef_4 (B.2.3)	0.1608	0.8515	pd6_coef
HR_robusttrend (B.1.4)	0.1546	1.1672	HRtrend
pd9_coef_1 (B.2.3)	0.1485	0.8620	pd6_coef
RRImax (B.1.7)	0.1474	1.1588	—

HRrange (B.1.7)	0.1425	1.1532	HR_range_ratio
Hist_RR_7 (B.1.15)	0.1386	0.8706	—
RR_uptick_ratio (B.1.12)	0.1345	1.1440	—
Number_comorbs_3 (B.2.5)	0.1313	0.8770	—
Hist_RR_6 (B.1.15)	0.1304	0.8777	—
QRStrend (B.1.4)	0.1094	0.8964	—
RR_transient_trend (B.1.10)	0.1082	1.1143	—
HasMultipleComorbs_0 (B.2.6)	0.1061	0.8993	—
RR_pval_ftest (B.1.8)	0.1011	1.1064	—
RRrangeratio (B.1.7)	0.0893	0.9146	—
Number_comorbs_2 (B.2.5)	0.0890	0.9148	—
Number_comorbs_1 (B.2.5)	0.0867	0.9170	—
HR_cubicfit_thirdcoef (B.1.9)	0.0863	1.0901	HR_cubicfit_secondcoef
RR_cubicfit_thirdcoef (B.1.9)	0.0835	1.0871	RR_cubicfit_secondcoef
QRS_b60min (B.1.4)	0.0740	1.0768	—
sdann (B.1.16)	0.0711	1.0737	—
QRSmad (B.1.7)	0.0693	0.9330	—
apen (B.1.18)	0.0670	0.9352	spen
Hist_RR_33 (B.1.15)	0.0656	0.9365	—
pd7_coef_0 (B.2.3)	0.0647	0.9373	—
HR_pval_ftest (B.1.8)	0.0579	1.0596	—
HR_catcoef_1 (B.2.2)	0.0519	1.0533	—

RRI_tri_area_sd (B.1.19)	0.0487	1.0499	—
Number_comorbs_5 (B.2.5)	0.0477	0.9534	—
RRI_peakfreq_lf (B.1.17)	0.0451	0.9559	—
Hist_RR_8 (B.1.15)	0.0427	1.0436	—
pd7_coef_3 (B.2.3)	0.0386	1.0394	—
Number_comorbs_4 (B.2.5)	0.0379	1.0386	—
pd9_coef_0 (B.2.3)	0.0374	0.9633	pd6_coef
RR_catcoef_2 (B.2.2)	0.0342	1.0348	—
RR_catcoef_1 (B.2.2)	0.0318	0.9687	—
Number_comorbs_0 (B.2.5)	0.0194	0.9808	—
Hist_HR_7 (B.1.15)	0.0187	0.9815	—
pd9_coef_3 (B.2.3)	0.0185	0.9817	pd6_coef
Hist_RR_28 (B.1.15)	0.0153	1.0154	—
HR_catcoef_2 (B.2.2)	0.0148	0.9853	—
Number_comorbs_8 (B.2.5)	0.0127	0.9874	—
RR_recentmad (B.1.7)	0.0102	1.0103	—
QRS_mww (B.1.8)	0.0087	0.9913	—
Number_comorbs_6 (B.2.5)	0.0081	0.9919	—
QRS_quadfit_firstcoef (B.1.9)	0.0074	0.9926	—
Number_comorbs_9 (B.2.5)	0.0067	0.9933	—

RRI_sd1sd2ratio (B.1.18)	0.0056	0.9944	—
Number_comorbs_7 (B.2.5)	0.0056	0.9944	—
pd9_coef_2 (B.2.3)	0.0041	0.9959	pd6_coef
HR_catcoef_3 (B.2.2)	0.0016	1.0016	—

HR - Heart Rate, RR - Respiration Rate, QRS - QRS complex amplitude, RRI - RR interval.

[a] Odds ratio represents the odds of predicting positive class in the presence of one unit increase in the respective feature, compared with the odds of predicting positive class in the absence of one unit increase in the respective feature [77].