Universidade do Minho
Escola de Engenharia

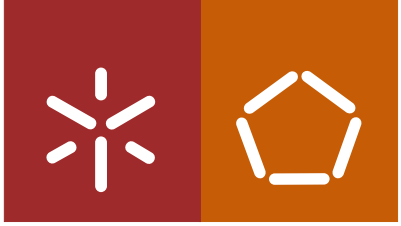
Cármina Augusta Pereira Azevedo

On the improvement of sleep onset latency detection and sleep/wake classification using cardiorespiratory features

July 2016

Cármina Augusta Pereira Azevedo **On the improvement of sleep onset latency detection and sleep/wake classification using cardiorespiratory features**

UMinho | 2016



Universidade do Minho
Escola de Engenharia

Cármina Augusta Pereira Azevedo

On the improvement of sleep onset latency detection and sleep/wake classification using cardiorespiratory features

Master dissertation in Medical Informatics
Integrated Master in Biomedical Engineering

Dissertation supervised by
Paulo Novais
Pedro Fonseca

July 2016

ACKNOWLEDGEMENTS

First of all, I want to express my appreciation to my research supervisors, Paulo Novais and Pedro Fonseca, for providing me the opportunity to take place in such an interesting research project and to experience the amazing work environment that Philips Research provides. In particular, I want to leave a special note to Pedro Fonseca, for pushing me out of my comfort zone, allowing me to develop further than what I initially imagined I could. Also, I am very grateful for the patience and systematic guidance provided throughout the internship. It was a pleasure to learn from such a good professional.

During my work at Philips Research, I was blessed with a dear group of colleagues. I appreciate their friendship and all the culture trading, which has allowed me to develop into a better person. This is particularly intended to Laura Secondulfo, Zihan Wang, Sreejita Gosh, Rick van Veen and Gabriele Papini. Dank yullie wel!

I am indebted to Ana Leitão and Luís Torres for their kindness, interest and useful advices before and during the internship, allowing me to properly prepare for important situations ahead.

I am profoundly thankful, from the bottom of my heart, to my treasure friends in Portugal, for their active presence in my decision making and help to go through difficult paths with a smile on my face. To Paulo Félix, Sara Rocha, Pedro Fernandes, Ana Assunção, Daniela Rijo, Raquel Ferreira, Janete Alves, João Rocha, Beatriz Ferreira, Alexandra Marques, Joana Oliveira, Ana Leal, Nuno Marques, and to my brother, Nuno Azevedo.

I want to leave a personal note to my group of colleagues during the first year of my Master studies at Universidade do Minho, namely Ana Quintas and Margarida Costa, who have fought the same fights as me to reach this point, and have become good friends of mine.

I am grateful my dear extended family, for the uncountable kind words and acts they have had towards me, not only during this particular phase of my life, but since I can remember. This is intended to my one-of-a-kind uncle and friend José Pereira, kind godfather Luís Pereira, and goodhearted aunt Arminda Pereira.

Finally, I want express my deep gratitude to my beloved parents for encouraging my academic journey since day one and raising me to always give my best. Without their support - both spiritual and material - I would not have come this far. I dedicate this work to them as a demonstration of my appreciation for the unconditional love and care.

ABSTRACT

This document describes an investigation performed at Philips Research (Eindhoven, The Netherlands) which aimed at improving the performance of Philips current *sleep/wake* classification methods using portable devices based on unobtrusive cardiorespiratory signal modalities. Particularly, this research focused on improving the detection of the Sleep Onset Latency (SOL) parameter.

Using a data set with recordings of healthy subjects, several alternative classification models were built, evaluated and compared to the current classifier.

It was found that the performance of the current classifier, regarding SOL detection, decreases with increasing SOL, leading to an underestimation of this parameter, and possibly undervaluation of symptoms of sleep disruptions or even sleep disorders, during medical diagnosis.

The main issue associated to this fault is that the current classifier is trained with examples from the entire night and therefore, for subjects with extended SOL periods, fails to capture the characteristics of *wake* before the initiation of sleep.

In this report a new method of distinguishing *sleep* from *wake*, to be applied with recordings of subjects with SOL over 30 minutes, is proposed. The new method comprises two steps: one specially dedicated to identify *wake* before the initiation of sleep (and more accurately detect the moment of Sleep Onset (SO)), and the other one to distinguish *wake* after SOL. Hence, it requires the use of two classifiers which differ regarding techniques for feature selection and are trained with examples of different periods of the night recordings.

KEYWORDS Machine learning, sleep staging, sleep/wake detection, sleep classification, sleep monitoring, sleep onset latency, cardiorespiratory features, actigraphy.

RESUMO

O presente documento descreve uma investigação desenvolvida na instituição Philips Research (Eindhoven, The Netherlands). O objetivo deste trabalho é melhorar o desempenho de atuais métodos Philips de classificação *sleep/wake* que utilizam dispositivos portáteis baseados na aquisição de sinais cardiorespiratórios. Em particular, este trabalho foca o melhoramento do desempenho desta tecnologia na deteção do período de latência de sono.
Utilizando um *dataset* que inclui registos de gravações noturnas de sujeitos saudáveis, vários modelos de classificação foram construídos, avaliados e comparados com o modelo atual.
Verificou-se que o desempenho do classificador atual, no que diz respeito à deteção do período de latência de sono, é inferior para sujeitos com dificuldade em adormecer (com latência de sono superior a 30 minutos [1]) o que conduz a uma subestimação deste parâmetro e, possivelmente, à subestimação de sintomas de distúrbios associados com o sono, aquando do diagnóstico médico.
Esta falha no desempenho está relacionada com o facto de o modelo de classificação atual ser treinado com exemplos de gravações noturnas completas, fazendo com que, para sujeitos com períodos de latência de sono prolongados, as caraterísticas da classe *wake* antes da iniciação do sono, não sejam bem capturadas.
Nesta dissertação é proposto um novo método para a deteção *sleep/wake* destinado a pessoas com latência de sono superior a 30 minutos. Este método inclui dois passos: o primeiro destinado especificamente à identificação da classe *wake* durante o período de latência de sono (detetando-o a sua duração com maior eficácia) e o segundo com o objectivo de distinguir *wake* durante o restante tempo da noite de sono. Assim, torna-se necessária a utilização de dois classificadores que diferem relativamente às técnicas utilizadas para a seleção de *features* e utilizam diferentes exemplos de treino, isto é, períodos distintos das gravações noturnas.

PALAVRAS-CHAVE Aprendizagem-máquina, classificação de sono, deteção *sleep/wake*, latência de sono, *features* cardiorrespiratorias, monitorização do sono, actigrafia.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A

AASM American Academy of Sleep Medicine.

ANS Autonomic Nervous System.

B

BMI Body Mass Index.

BP Blood Pressure.

C

CFS Correlation Feature Selection.

D

DFA Detrended Fluctuation Analysis.

DS Deep Sleep.

E

ECG Electrocardiography.

EEG Electroencephalogram.

EMG Electromiography.

EOG Electrooculography.

F

FN False Negative.

FP False Positive.

FPR False Positive Rate.

FR Feature Restriction.

FS Feature Selection.

G

GT Ground Truth.

H

HF High Frequency.

HR Heart Rate.

I

IQR Interquartile Range.

L

LD Linear Discriminant.

LF Low Frequency.

N

NREM Non-Rapid Eye Movement.

P

PPG Photoplethysmogram.

PPV Positive Predictive Value.

PSD Power Spectral Density.

PSG Polysomnography.

PSQI Pittsburg Sleep Quality Index.

R

R&K Rechtschaffen and Kales.

REM Rapid Eye Movement.

RIP Respiratory Inductance Plethysmography.

S

SE Sleep Eficiency.

SO Sleep Onset.

SOL Sleep Onset Latency.

SWS Slow Wave Sleep.

T

TIB Time In Bed.

TN True Negative.

TNR True Negative Rate.

TP True Positive.

TPR True Positive Rate.

U

.

V

VLF Very Low Frequency.

W

WASM World Association of Sleep Medicine.

1

INTRODUCTION

Sleep is a fundamental process that we perform with daily regularity. Even though the average human spends a third of its existence sleeping, very few people have basic knowledge of how sleep works. In fact, even though sleeping and dreaming have been subject of interest and inquiry since the time of the ancient Greek philosophers, it was not until the decade of 1920 that sleep was studied in terms of scientific understanding [2].

Nevertheless, from experience, and even without any scientific support, we know that a night well slept has the power to make us feel refreshed and ready to take on the world for the day ahead. On the other hand, the absence of sufficient hours of sleep can evoke terrible feelings. Based on this we can speculate on the importance of sleep: not only do we spend a big slice of our lives sleeping, but also, the quality of our sleep highly impacts the remaining time (two thirds) of our lives. This brings us to another paradoxical, however very realistic, panorama: regardless of sleep being a natural process, more than half of the adult population claims experiencing some kind of trouble or difficulty related to sleeping [3].

In the modern day, we find ourselves living as part of culture of jet lag, global travel, 24-hour cable TV, Internet addiction and shift working. Despite the advantages and infinite possibilities that our current way of living might bring, it is dramatically affecting the way we sleep. More and more people are currently suffering from sleep disorders: the World Association of Sleep Medicine (WASM) [4] has pointed out that, nowadays, up to 45% of the world population suffers from sleep-related pathologies, such as insomnia, obstructive sleep apnea, restless legs syndrome and sleep deprivation in general. These are extremely worrying numbers which situate sleep disorders among the most common illnesses of our time, even though sleep is one of the main cornerstones of good health, along with diet and exercise.

Polysomnography (PSG) is the *gold standard* to perform sleep detection over the length of a night of sleep [5, 6]. This is extremely important since the diagnose and treatment of sleep disorders implicate the analysis of the sleep patterns of the patient, which are accurately provided by this method. Even so, despite offering accurate physiological measurements during sleep, PSG is based on manual sleep staging, which is time-consuming, and involves high costs regarding laboratory facilities, equipment and qualified team. Automatic sleep

stage classification, on the other hand, removes the human element from the equation and allows real-time sleep staging (useful for intervention studies) and remote monitoring (in-home sleep studies).
The electroencephalogram (EEG), as a method to measure sleep stages, has the major inconvenience of being uncomfortable for the patient and, therefore, obtrusive of sleep. Also, it requires the correct positioning of the sensors, performed by an expert.
These drawbacks have been motivating research regarding unobtrusive methods, such as sleep detection based on body movements (actigraphy) and cardiorespiratory activity [7]. Nevertheless, the performance associated with sleep staging using this technologies is still not as high as it would be desirable.

1.1 problem description

One of the parameters of interest derived from sleep stages is the Sleep Onset Latency (SOL) period [8], which measures the amount of time elapsed from the instant when a person lies in bed with the intention to sleep until the moment when that person eventually falls asleep. The SOL is a valuable indicator in the evaluation of sleep quality, sleep complaints and sleep disorders, such as insomnia and circadian rhythm disorders.
The problem with current unobtrusive approaches for *sleep/wake* classification is that they heavily rely on the idea that *wake* states usually coincide with periods of longer and/or more intense body movements. Actigraphy devices such as the Philips Actiwatch [9] actually rely on the detection of body movements, measured with a wrist-worn accelerometer, to automatically distinguish *wake* from *sleep*. While this observation is certainly true for the majority of the night - brief periods of awakening during sleep often occur together with body movements ('tossing and turning' in bed) - this is not always valid at the beginning of the night, when one is trying to fall asleep. In fact, the longer it takes for a person to initiate sleep, the worse the performance of actigraphy-based *sleep/wake* classifiers typically is, in particular in the detection of SOL [10]. This is easily explained, given that, usually, a person will try to lie as still as possible during SOL. This is particularly common in subjects who frequently experience long SOL periods. As a result, actigraphy-based SOL estimators will very often underestimate this measure and overestimate *sleep*, potentially leading to an underestimation of symptoms associated with sleep disruptions or even sleep disorders.
Even though introducing cardiorespiratory information has shown to improve the accuracy of *sleep/wake* detection [11, 7, 12], it has not completely solved the problem of the underestimation of the SOL parameter. The main issue is that these classifiers are trained for this task with examples from the entire night and, therefore, fail to capture the particular characteristics of *wake* before SOL.

1.2 OBJECTIVE

This work focuses on the improvement of the identification of SOL and overall *sleep/wake* detection based on unobtrusive cardiorespiratory signal modalities. It describes the methods that proved to be most suitable for the task, as well as their application and evaluation. An increased reliability of the classification output consequently leads to the improvement of the quality of unobtrusive diagnosis/treatment equipments available for home monitoring and to sleep professionals. Hence, improving the effectiveness rate of the treatments of sleep disorders, and, hopefully, contributing to an healthier society.

1.3 SOLUTION APPROACH

A new classification scheme is proposed, with the intent of overcoming the limitations of the current classifier and achieving a more correct estimation of *sleep/wake* and SOL detection. The proposed approach includes the usage of two distinct classifiers, to be applied on subjects with delayed SOL. The first classifier is designed with the purpose of recognizing *wake* in the beginning of the recordings, and the other to recognize *wake* throughout the night. This scheme allows to compute an estimation of SOL from the scoring provided by the first classifier, and apply the second one from that moment on.
All tests performed in order investigate the topic and all results presented in this report were obtained by the development of computational routines/algorithms, using high-level programming language and the environment MATLAB®.

1.4 ORGANIZATION

The following chapter is entitled 'Background' and includes a theoretical description on sleep related knowledge, described in literature, which is indispensable prior to the reading of the remaining contents of this report. Chapter 3, 'Sleep and Wake Classification', includes two sections. The first one, section 3.1 describes the guidelines followed to perform *sleep/wake* detection, in this work. In section 3.2, 'Current Performance Assessment', the performance of the current classifier is evaluated.
Chapter 4, 'Experimental Study on the Performance of SOL Detection', describes an investigation on how certain aspects of the classification process influence the results on SOL detection, specifically: time of the analysis; usual quality of sleep; feature transformation methodologies and feature selection/restriction. It is organized in three sections: data set and methods; results and discussion. This is also the adopted structure for the following chapters. On chapter 5, 'Impact of Actigraphy on the Performance of SOL Detection', we evaluate the influence of the actigraphy feature (and correlated features) on the per-

formance of SOL detection. Finally, on chapter 6, based on the discussions/conclusions obtained from the studies described on previously referenced chapters, a new method for *sleep/wake* and SOL detection is proposed and its performance is evaluated, in comparison to currently used classifiers. Chapter 7 concludes the report with a summary of the achieved results of interest and suggestions for possible directions of feature research.

2

BACKGROUND

The purpose of this chapter is to introduce the main concepts of human sleep physiology, architecture and mechanisms, which are critical to understand the approaches conducted in this research.

2.1 SLEEP HOMEOSTASIS

In the early 1980's, Alexander Borbéli defined sleep homeostasis as the *regulated balance between sleep and waking* [13]. He presented the theory of the **two-model process of sleep regulation** [13] which suggests that complex interactions between two independent physiological processes underlie sleep regulation. Those processes are:

- Process S, the **homeostatic process** (or sleep drive), which cumulatively increases during waking and decreases exponentially during sleep [13]. In other words, the longer the time we spend awake, the greater the pressure we will feel to go to sleep. By sleeping, sleep pressure is reduced and, gradually, propensity to wakefulness rises [14].
- Process C, the **circadian process**, is a sleep-independent mechanism that operates in cycles of approximately 24 hours, as a function of our biological clock, driven by the natural light-dark cycle of the day. The circadian process regulates distinct physiological rhythms, such as body temperature and hormone secretion (for example, melatonin and cortisol), which influence sleep [15]. Hence, by determining the alternation between cycles of high and low sleep propensity, process C is related to the regulation of sleep timing [14].

The two-process model interaction can be visualized graphically on figure 1.

As illustrated in figure 1, during a regular day of wakefulness, process S continuously builds up, inducing sleep drive. Simultaneously, process C increases its awareness effect, opposing to the latter. It is not until the alerting signal drops (usually during the evening), that sleep pressure becomes dominating enough to allow sleep to be initiated.
There are several factors that can impact the balance of our sleep-wake system [16]. For

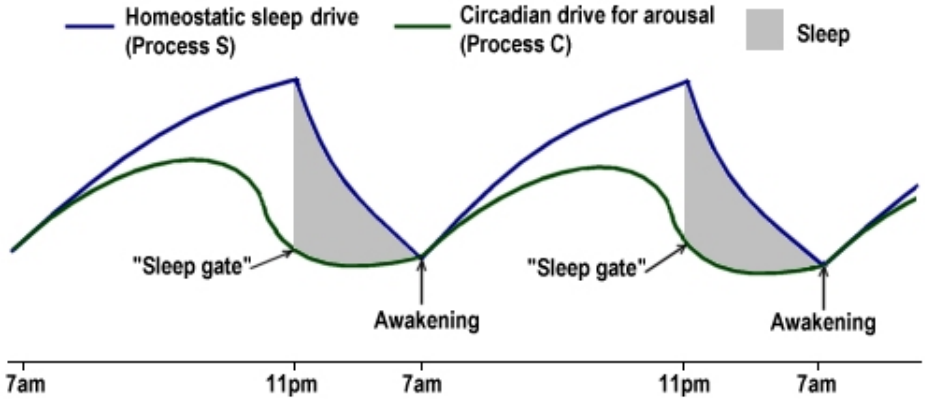


Figure 1.: The two-model process of sleep regulation: synchronized interaction between process S and C [2].

example, genetic information, stress and health conditions (such as pain and anxiety) can affect the initiation of sleep [2]. Also external factors, such as light exposure, arousal (noise), medication, caffeine and alcohol intake, among others, generally tend to increase the upper-threshold by which we fall asleep. Thus, due to the dis-synchronization between the environment and our internal biological clock, there is a fluctuation which results in a feeling of fatigue. This is why fatigue is so associated with jet lag and shift working [17].
On a cellular level, the transitioning process between the conditions of wake and sleep is controlled by brain cell interaction: some neurons are responsible for promoting wakefulness while others induce sleep [16]. When the alerting parts of the brain are more active they inhibit the sleep-promoting brain areas. Consequently, the state of wakefulness is present. Analogously, when the sleep-promoting areas are dominantly active, they have an inhibitory effect on the areas responsible for promoting wakefulness, resulting in sleep.
This mutual inhibition can be explained by a **flip-flop switch** model [16], as in an electrical circuit, that either leads to the state of wakefulness or sleep, in rapid transitions. Typically, the transitions between the two states occur when the mechanism is triggered by factors related to the homeostatic sleep drive and the circadian rhythm, as explained above.

2.2 SLEEP ARCHITECTURE

While sleeping, our brain exhibits distinct electrical activity, which can be measured by an Electroencephalogram (EEG) and used to characterize the process of sleep and to identify sleep stages [16].
The transition between stages occurs in a cyclic pattern which alternates between Rapid Eye Movement (REM) and Non-Rapid Eye Movement (NREM) sleep [18, 19]. The former occupies about 20 to 25% of the night, while the latter accounts for the remaining percentage. For the average healthy adult, during a typical night of sleep, approximately every 90

minutes (min), the cycle repeats itself. One cycle can be discretized by order of stage occurrence as: $N1 \rightarrow N2 \rightarrow N3 \rightarrow N2 \rightarrow REM$, where N1, N2 and N3 are, respectively, the NREM stages 1, 2 and 3, which compose NREM sleep, according to the AASM [18, 19]. From 1968 to 2007, prior to establishment of the AASM rules [20], the Rechtschaffen and Kales (R&K) [21] scoring was used, and it further divided the N3 stage into stages III and IV.

Each stage will now be described, from an electroencephalographic point of view. The description is accompanied by figure 2, which reveals the typical EEG recording form for each stage.

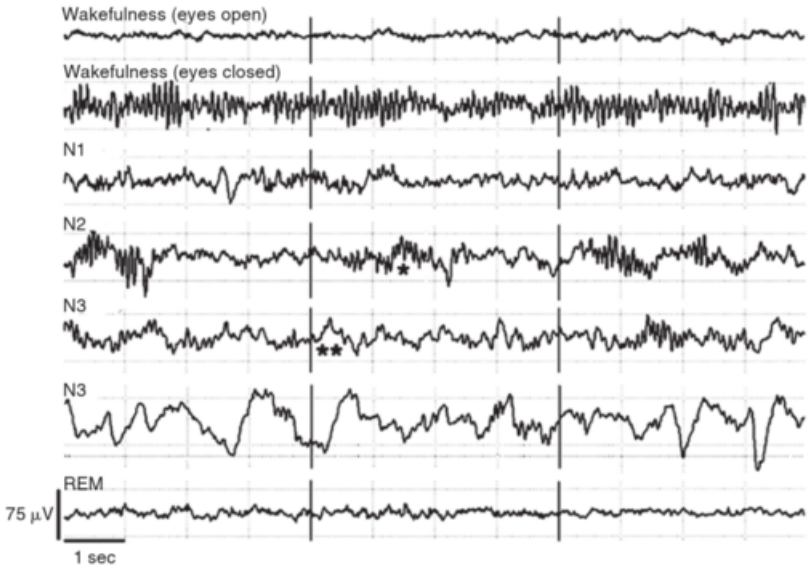


Figure 2.: Human EEG recordings during wakefulness and sleep stages (*sleep spindles; **slow wave) [22]

During **wakefulness**, when there is a state of alertness and active mental concentration, the EEG form verified is the β-wave, which has the highest frequency and the lowest amplitude when compared to other sleep stages [23]. However, when a person is awake but with eyes closed, with the intention of falling asleep, as it is the case during the SOL period, the EEG records α-waves [23].

N1, often referred to as light sleep, is the first stage of sleep experienced in the cycle. It describes the transition from wakefulness (SOL period) to drowsy sleep. Hence, the EEG recordings display brain waves shifting from α-band activity to Θ-band activity (still accompanied by brief bursts of α-activity) [23, 2]. This combination of brain activity often gives people the sensation of still being awake, when, in fact, they might already be experiencing N1 sleep. The length of this stage over an average night of sleep corresponds only to 2 to 5% of total sleep time.

The following stage, **N2**, still falls into the category of light sleep. During N2, the EEG

recordings reveal irregularities in the brain wave pattern, namely, sleep spindles (bursts of oscillatory brain activity) and K-complexes [23]. Both wave forms are illustrated, closely, in figure 3.
In comparison to the previous stage, N1, conscious awareness completely disappears [18] .

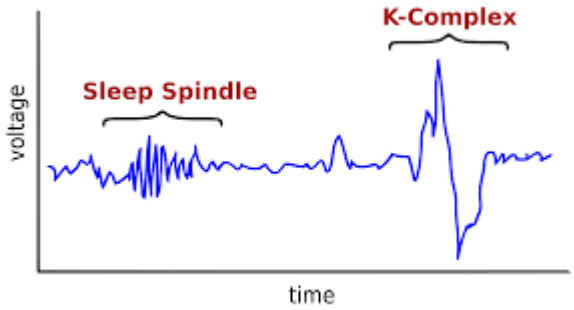


Figure 3.: EEG typical signal of the N2 stage, containing a sleep spindle, on the left, and a K-complex, on the right [16].

The succeeding stage, **N3**, is the deepest and most restorative stage of sleep, therefore it is known as Deep Sleep (DS) [2]. The EEG recordings reveal synchronized, low frequency and high amplitude Δ-wave activity, hence, DS is also known for the designation of Slow Wave Sleep (SWS) [19]. During this stage, people will hardly respond to environmental stimulus and there is little (or none) muscle activity. Parasomnias, such as sleep-walking, sleep-talking and night terrors, occur during this stage [23]. It accounts for 15% to 20% of total sleep time.
Finally, the REM stage, often called paradoxical sleep [18], is characterized by a random and rapid movement of the eyes and simultaneous paralyzation of most of the body muscles [18, 19]. Despite the almost absence of movements, from an EEG point of view, REM resembles wakefulness because of the presence of desynchronized, low-amplitude and high-frequency β-waves [18, 19]. This fact makes REM sleep even more intriguing. This stage is also characterized by the ability to dream vividly; heart rate variability; irregular breathing; and Electromiography (EMG) activity [24].
The graphical representation of the sleep stages experienced by a person during the course of one night is entitled **hypnogram** [18]. On figure 4 there is an example of an hypnogram of an healthy adult.

As the pattern repetition proceeds, it is verified that N3 tends to diminish its length time, from each cycle to the next, while the opposite happens to REM sleep [18, 19], as observed on figure 4.

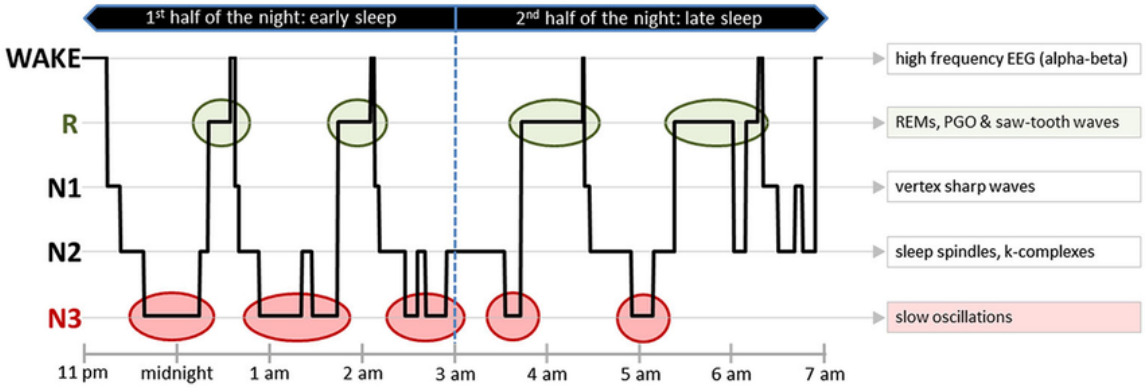


Figure 4.: Example hypnogram of an healthy adult human over 8-hour sleep [25]. On the left side of the picture sleep stages are labeled according to the AASM rules. R stands for REM sleep. On the right side of the figure there is a description of typical predominant EEG activity associated to each stage.

2.3 POLYSOMNOGRAPHY

As mentioned on chapter 1, the *gold standard* technique to measure sleep is through PSG. It comprises monitoring of brain function, using EEG; eye movements, using Electrooculography (EOG); heart rhythm, using Electrocardiography (ECG); body movements and skeletal muscle tone, using EMG; respiratory airflow, using nasal-oral thermistors and nasal pressure transducers; respiratory effort, using belts; and CO_2 saturation, using pulse oximetry [26]. Figure 5 represents the montage of a standard PSG and figure 6 allows for a closer look into the positioning of some of the sensors, specifically those located on the face/head of the patient.

PSG recordings are performed in a sleep laboratory and the overnight sleep stages are manually scored on a 30 second epoch basis, by trained sleep experts [26, 21].
Even though the PSG provides a complete and reliable set of measurements that are of interest for sleep stage detection, the equipment is so uncomfortable, that it tends to be disruptive of sleep. Hence, a less obtrusive method is desired.
As mentioned in section 1.2 and suggested by the title of this work, the classification framework applied in this research makes use of only cardiorespiratory signals for the purpose of sleep stage detection, therefore being less unpleasant for the patient, and consequently less sleep disruptive of sleep.
The following section provides a description on how cardiovascular and respiratory patterns vary through sleep stages.

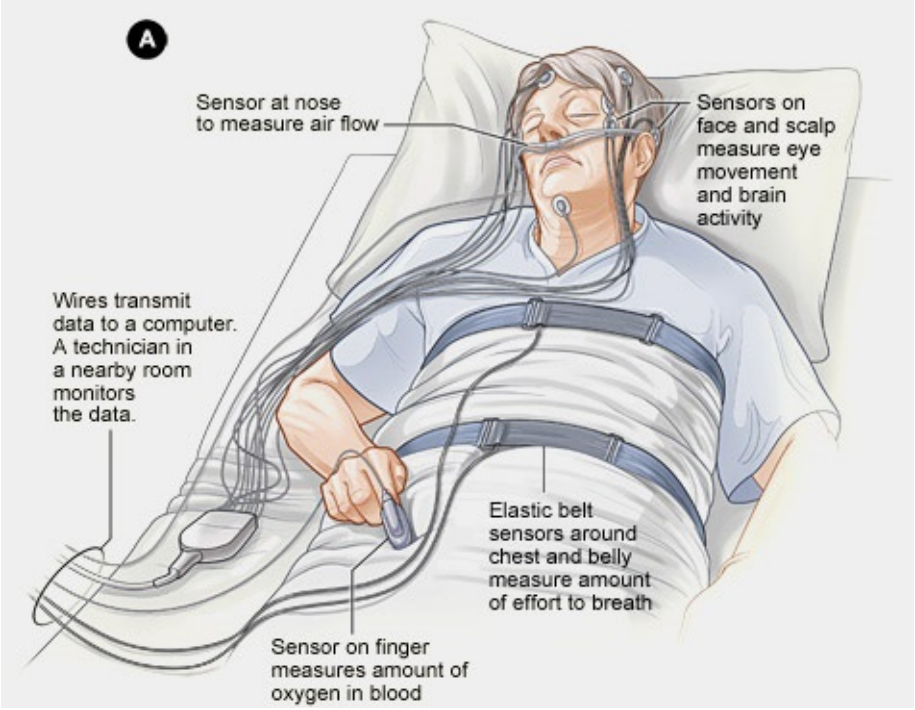


Figure 5.: Drawing representing a standard PSG montage on a patient [27].

2.4 CARDIORESPIRATORY-BASED SLEEP STAGE CLASSIFICATION

Cardiorespiratory activity is characterized differently on each sleep stage, hence, it is a good measure of sleep staging. The differences registered on cardiorespiratory signals during sleep are due to manifestations of the Autonomic Nervous System (ANS), which includes in its structure sympathetic and parasympathetic (or vagal) activity. Between these two tones there is a balance (the sympathovagal balance) on which one activates an action while the other suppresses it [24]. For instance, a decrease on Heart Rate (HR) variability is usually linked to vagal activity, while an increase is, most likely, associated with the sympathetic nerves [24].
After SOL, vagal activity increases, while sympathetic activity decreases [24]. Vagal predominance during sleep is more accentuated during NREM stages. Therefore, as we progress from wakefulness to NREM sleep there is a noticeable decrease in HR and Blood Pressure (BP), which gradually become more regular, reaching their smoothest state during N3 sleep. During periods of REM and brief awakenings, both BP and HR increase and become more variable [16].

When it comes to breathing, when we are awake, it is affected by several factors, such as: speech, emotions, exercise and posture. When we enter NREM sleep our breathing rate (BR) starts decreasing and becomes more regular, both in amplitude and frequency [30]. Also, as we progress from lighter to deeper sleep, our mean inspiratory flow decreases. On

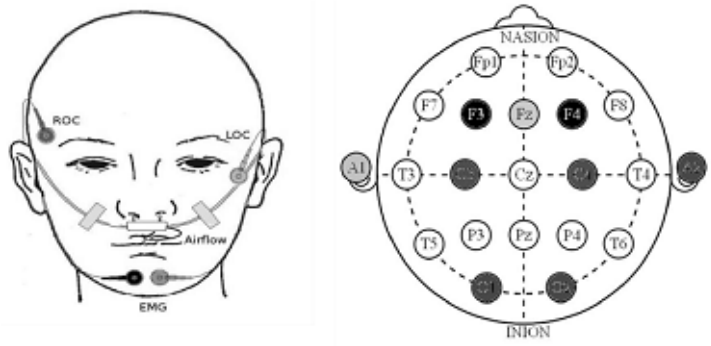


Figure 6.: Example of a standard PSG montage sensors that are located on the head of the patient. On the panel on the left: right outer canthus and left outer canthus which, together, constitute the EOG; airflow sensors and submental EMG. At the right panel: EEG with the essential electrodes [28].

the other hand, during REM sleep, BR is irregular, with sudden changes both in amplitude and frequency, correspondent to the burst of the movement of the eyes [16].

Figure 7 displays the hypnogram of an healthy subject, with correspondent activity counts and variations on BR and HR. Regarding BR and HR, there is an high correlation with the hypnogram. Regarding activity, after SOL, movements are almost absent during the course of the night, except for spontaneous peaks correlated to brief awakenings.

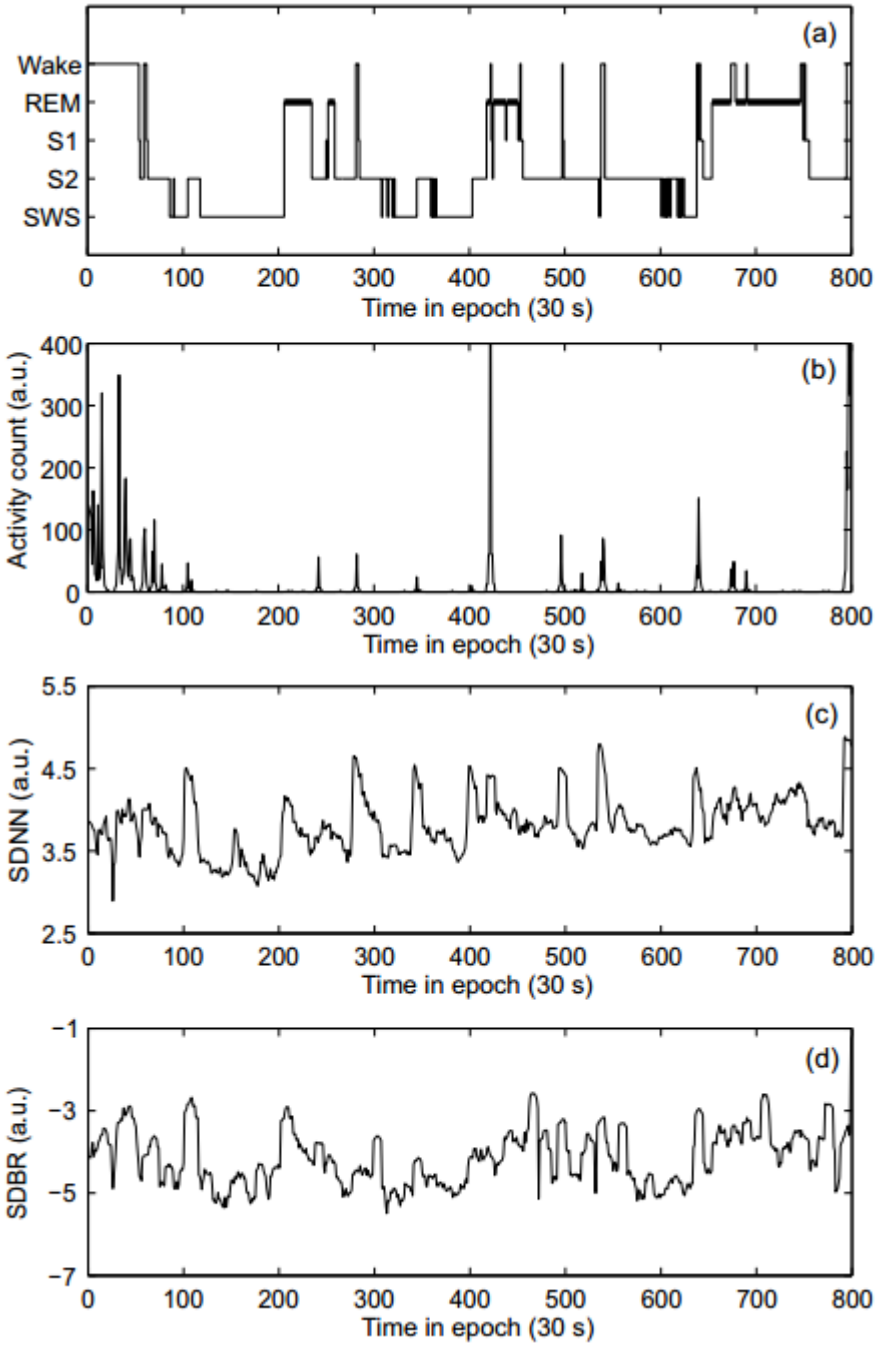


Figure 7.: Example of sleep recordings of an healthy adult during the course of one night: (a) hypnogram (S1 and S2 represent sleep stages N1 and N2); (b) movement counts; (c) standard deviation of normal-to-normal heartbeat intervals (SDNN); (d) standard deviation of breathing rates (SBDR) [29].

3

SLEEP AND WAKE CLASSIFICATION

This chapter includes a description of the data utilized in this research and of the machine learning and statistic procedures applied. Furthermore, it addresses the status (performance-wise) of current classification methodologies available at the beginning of this work, for the task of *sleep/wake* detection.

3.1 DATA SET AND METHODS

Sleep scoring is performed using supervised learning algorithms, which infer a classification output, based on examples of annotated training data. Accordingly, the data includes a set of important measurements which are assigned to a correspondent class (the label/annotation).

3.1.1 *Data Set*

In this research, a data set comprising of night-recordings from 180 subjects, was used. Table 1 presents basic demographic information, such as age, gender and body mass index (BMI), and also sleep measurements acquired from the recordings, namely time spent in bed (TIB), as the amount of time elapsed from the moment of lights off until the moment of lights on; SOL time and sleep efficiency (SE), which is the proportion of sleep, during TIB, computed according to equation 1.

$$SE(\%) = \frac{time\ asleep}{TIB} \times 100; \tag{1}$$

This is an adult healthy population, with a Pittsburg Sleep Quality Index (PSQI) [31] scored less than 6, that fits several criteria such as the absence of: sleep complaints; previous diagnosis of sleep disorders; shift work and/or depressive symptoms.
On average, this is a population that sleeps the consensual number of hours, with reasonable SOL time, and relatively high SE [32]. Nevertheless, the standard deviation associated particularly to SOL time is high, and so is the maximum value encountered for it. This

Table 1.: Summary of subjects demographic information and sleep measurements

Parameter	$\overline{x} \pm \sigma$	Range
N	339 recordings (180 subjects)	
Sex	98 females (54.4%)	
Age (year)	50.10±19.68	$20 - 95$
BMI (kg/m^2)	24.55±3.49	16.98-35.25
TIB (hour)	7.85±0.62	2.45-9.33
SOL (min)	23.47±21.50	1.50-141.00
SE (%)	80.76±11.76	21.39-97.89

indicates that, even though the mean values are considered ordinary and non-alerting [33], there are subjects among the population, that have experienced troubles in falling asleep, at the nights of the recordings. Considering that this an healthy population, the extended SOL times are probably due to the *first night effect* described in literature [34].
The information presented on table 1 refers to the grouping of recordings from 3 different data sets, namely: Boston and Eindhoven data sets, and part of the data set acquired in the SIESTA project [35].

Boston and Eindhoven data sets

The Boston and Eindhoven data sets include single-night PSG recordings and synchronized actigraphy recordings (acquired with a Philips Actiwatch [9]) of 15 healthy subjects (considering the criteria mentioned above, in section 3.1.1). Nine subjects were monitored (Alice 5 PSG, Philips Respironics) in Boston, USA, during the year of 2009, at the Sleep Health Center, while the remaining 6 were measured (Vitaport 3 PSG, TEMEC) in Eindhoven, the Netherlands, during 2010, at the High Tech Campus. Each subject provided an informed consent and the study protocol was approved by the Ethics Committee of the two sleep laboratories.
The PSG recordings are comprised of multi-channel signal modalities such as EEG channels recommended by the AASM [20]; EMG; EOG; a 2-lead ECG; oxygen saturation; and thoracic respiratory effort.
Sleep stages were manually scored, on a 30-s epoch basis, by sleep experts, according to the AASM guidelines [20] as *wake*, *REM* sleep and *N1-N3*, for NREM sleep.
Demographic information considering subjects from the two data sets are presented on tables 2 and 3.

SIESTA regular data set

The SIESTA project [35] included sleep monitoring in seven different sleep centers, in five European locations, over a period of 3 years from 1997 to 2000. The SIESTA regular data set

Table 2.: Demographics for subjects on the Boston data set.

Parameter	$\overline{x} \pm \sigma$	**Range**
N	9 recordings (9 subjects)	
Sex	8 females (88.9%)	
Age (year)	31.89 ± 12.82	$24 - 58$
BMI (kg/m^2)	25.31 ± 3.81	$21.00 - 31.20$
TIB (hour)	7.09 ± 0.48	$6.56 - 7.69$
SE (%)	91.50 ± 3.75	$86.02 - 97.89$
SOL (min)	14.72 ± 10.56	$3.50 - 33.00$

Table 3.: Demographics for subjects on the Eindhoven data set.

Parameter	$\overline{x} \pm \sigma$	**Range**
N	6 recordings (6 subjects)	
Sex	2 females (33.3%)	
Age (year)	29.67 ± 5.92	$23 - 36$
BMI (kg/m^2)	22.98 ± 2.06	$20.24 - 26.54$
TIB (hour)	6.55 ± 2.49	$2.45 - 8.78$
SE (%)	92.33 ± 4.57	$83.85 - 95.87$
SOL (min)	9.83 ± 4.83	$4.00 - 16.50$

includes only night recordings of healthy subjects, that fit the criteria mentioned in section 3.1.1. Most subjects spent two consecutive nights in the sleep laboratories. Sleep scoring was manually performed by sleep clinicians, based on all PSG channels, also on a 30-s epoch basis, according to the R&K rules [21]: *wake*, *REM*, and *S1-S4* for NREM sleep.
Table 4 presents demographic information regarding the subjects from the SIESTA regular data set, whose recordings were utilized in this research.

Table 4.: Demographics for subjects on the SIESTA regular data set.

Parameter	$\overline{x} \pm \sigma$	**Range**
N	324 recordings (165 subjects)	
Sex	88 females (53.3%)	
Age (year)	51.84 ± 19.42	$20 - 95$
BMI (kg/m^2)	24.56 ± 3.52	$16.98 - 35.25$
TIB (hour)	7.90 ± 0.50	$5.32 - 9.33$
SE (%)	80.24 ± 11.75	$21.39 - 97.54$
SOL (min)	23.97 ± 21.79	$1.50 - 141.00$

3.1.2 *Sleep Scoring*

All stages regarding *sleep*, either REM sleep and N1-N3 (or S1-S4) for NREM, were merged into a single class, labeled as *sleep*, and were labeled *wake* otherwise, since the purpose of this research is to distinguish sleep from wakefulness and not necessarily to distinguish between different stages of sleep. Thus, the output which results from classification, on a 30 seconds epoch basis, is binary. *Wake* was identified as the positive class and *sleep* as the negative class.
In this work, SOL was recognized as the first epoch within the occurrence of three consecutive sleep epochs (regardless of their sleep stage correspondence).

3.1.3 *Framework for Sleep Stage Classification*

This section presents a brief explanation of the steps involved in the automatic process of sleep classification, which was not developed during this work or by this author, but was, anyhow, utilized in this research.

Data Acquisition & Feature Extraction

As mentioned previously in section 3.1.1, the data was acquired from PSG data collection and actigraphy recordings. However, since the raw recordings are not enough to precisely characterize sleep, there is the need for a process of extracting more relevant characteristics from the recordings. Such a characteristic is named **feature** and such a process is referred to as **feature extraction**.
The **feature set** (collection of extracted features) used in this work comprises a total of 169 features (which have all been previously described in literature and applied for sleep staging in healthy subjects [36]), expressing information about the cardiac system; the respiratory system; and the coupling interaction of both systems. There is an additional feature which describes estimated actigraphy information from the body movement artifacts present in ECG and respiratory inductance plethysmography (RIP) [37], given that the SIESTA data set does not include actigraphy recordings.

CARDIAC FEATURES - Many based on statistics computed over R-R intervals calculated from the QRS complexes of ECG recordings. The QRS complex is the designation of three typical waveforms of an ECG, as shown on figure 8, and the R-R interval is the time elapsed between two consecutive R waves [38].

Some features express the average HR per epoch; the n^{th} percentile; the standard deviation; and the range of the beat interval lengths. Other features are computed from

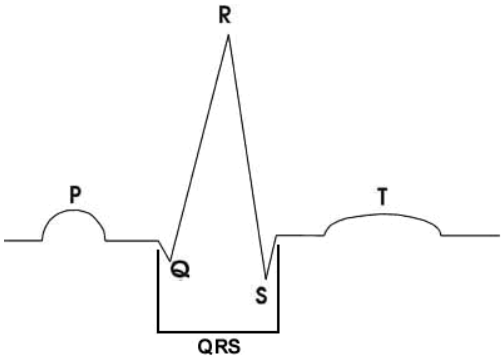


Figure 8.: Typical representation of a QRS complex of an ECG recording.

the power spectral density (PSD) analysis, regularly over three different frequency bands: very low frequency (VLF), $0.005 - 0.04$ Hz, low frequency (LF), $0.04 - 0.15$ Hz, and high frequency (HF), $0.15 - 0.45$ Hz, and from the modulus and the phase of the pole in the HF band. Other features describe the regularity of the signal over different time scales. For instance, detrended fluctuation analysis (DFA) is used to identify longer-term correlations in the signal, and sample entropy to quantify the self-similarity of the signal over a given time period [7, 11, 12, 39, 40].

RESPIRATORY FEATURES - Based, mainly, on respiratory effort measured, for example with a RIP belt around the thorax or around the abdomen. Several properties of respiration rate and amplitude have shown to be linked to different sleep stages [11, 41, 42]. Also, self-similarity measurements using dynamic warping have been described as useful for detecting wake states [43].

CARDIORESPIRATORY COUPLING FEATURES - Expressing the phase synchronization between R-R intervals from ECG or beats from a photoplethysmogram (PPG), and the respiratory phase measured from RIP or from PPG during a number of breathing cycles [44, 45, 46, 47].

Feature Transformation

Following the feature extraction process, the features can be submitted to transformations in order to make them more optimized and adapted to the requirements of the classifiers. For better results, the transformation is done in two steps: normalization and post-processing.

FEATURE NORMALIZATION - Aims at reducing *between-subject variability*, i.e. features that significantly vary between subjects, due to physiological or equipment-related variations. This is useful since the classifier expects a generalized model and, not having it would,

most likely, lead to a decrease in the accuracy of the classification. By normalizing features, they are brought to a common base line, making them more easily comparable between subjects.
There are different methods to normalize features which are included in the framework. However, they will not be described here, since exploring that is not the goal of this research.

FEATURE POST-PROCESSING - Allows the elimination of undesired components in the features, such as long-term trends. This is extremely useful since these are components that can diminish the discriminative power of a feature over time. Again, there are several post-processing methods that can be applied, still, they will not be described in this work.

For each feature the best sets of normalizing and post-processing methods are chosen. This is a complex task which was not developed during this work and therefore will not be described in this document.

Feature Selection

Feature Selection (FS) consists in evaluating and selecting the most appropriate features for the classification task, with the goal to achieve higher classification performance. It allows faster computational time of the remaining steps to the classification process, and reduces its complexity.
The evaluation is based on attributes such as relevance, discriminatory power, redundancy and correlation between features.
Since the feature set available for this work is very wide, most likely, it will comprise redundant features. A feature is considered redundant if there is one or more features besides it that share its (or similar) information. By being highly correlated to other features, redundant features add no information to the classification. Moreover, by including more features than necessary in the training process it is likely that there will a problem of **overfitting** which will lower the capacity of the classifier on generalizing well to new data [48]. Therefore, redundant features should be removed from further participating in the analysis. As a result, at the end of this phase, from the complete feature set, there should be only a subset of features selected to proceed with in the classification.
In this work, the process of FS is performed by a Correlation Feature Selection (CFS) algorithm [49], which is based on a greedy forward search. This method starts with an empty set of features which is progressively adapted by greedily including the next *best* (most relevant) feature, according to an heuristic metric evaluation of its discriminative power and (lack of) correlation with already added features. The process stops when a desired amount of features is reached.

Classification

Classification is the process of assigning each epoch to a specific class, based on the characteristics of that epoch, which are expressed by a feature vector. In this step, there are two desired goals: building a model that 1) provides the best possible fit and that 2) is robust against variability between subjects and performs well on unseen data.
As already referred, this work aims at investigating the binary problem of the identification of each epoch as *sleep* or *wake*. For this classification problem a Bayesian Linear Discriminant (LD) classifier [50] was used. Previous sleep research has proved that the LD classifier is the most adequate for the task of *sleep/wake* detection [7]. The basic idea behind the LD algorithm, introduced by Fisher, is to find a linear discriminant that yields a maximum separation between two classes [50].
Classification requires a **training** phase, so that the classifier can be designed and modeled to available example data, and a **testing** phase, to estimate the error rate of the trained classifier. Training and testing samples must be different and statistically independent, in order to get reliable predictions in future classification [51].
Over time, more than one way of combining training and testing samples for *error estimation* has been proposed:

HOLD-OUT [51] - this method suggests that the available data should be split into two disjoint subsets, as exemplified on figure 9, for a single train-test experiment. One group is used for the task of training including FS while the other is held for the task of testing.

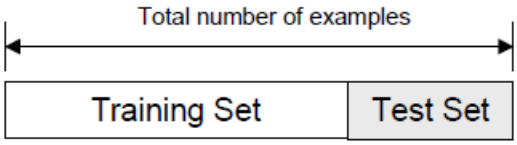


Figure 9.: Exemplified distributions of examples as training and testing subsets.

One limitation of this method is that it is not appropriate if the data set available is small. Generally, in those conditions, the testing set is valuable for training and, by having set it aside, the performance of the prediction will be compromised, leading to biased results. Notwithstanding, in the case of not so small data sets, data will still be wasted.
Another drawback of this method is that, since the results are so dependent on the choice of the training/testing set of samples, they will be misleading in the event of an unfortunate division. For instance, it might be too easy/difficult to classify certain examples of data in the testing set, leading, again, to biased results.

CROSS-VALIDATION [52] - an alternative method that overcomes the difficulties associated to the hold-out method, at the cost of more computations. It still consists in performing the training and testing on separate subsets of data, but the process is repeated several times.

One type of cross-validation is the **K-fold cross-validation**, according to which the data is partitioned in K bins of approximately equal size and the model is trained and validated K times in a loop. Figure 10 exemplifies a splitting of the data and training/testing according to this model.

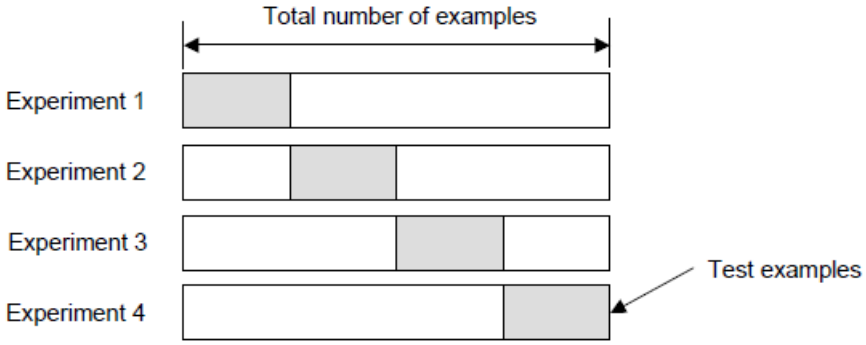


Figure 10.: Sample K-fold cross-validation with K=4.

The idea is that, on each experiment, $K-1$ folds are used for training and the remaining one for testing. FS occurs on the data that is not left out, during each cross-validation loop.

The main advantage of this method is that, at the end of the process, all examples have been used for both training and testing. The overall performance is computed based on the average performance achieved in each fold. Hence, the result is less sensitive to the partitioning of the data set and more accurate.

3.1.4 *Performance Assessment*

On the next chapters of this report, several performance metrics will be referenced. Some will be related to SOL-detection performance while others to the overall epoch-by-epoch sleep scoring. To facilitate and avoid repetitions, all measures are listed and briefly explained below.

- SOL-detection performance measures:

L_1 **ERROR (MIN) -** Taxicab or Manhattan distance, which represents the length between two vectors p and q in an n-dimensional real vector space:

$$L_1(p,q) = \sum_{i=1}^{N} |p_i - q_i| \tag{2}$$

In this work, this metric is applied to the vectors of the measurements of the detected SOL moment (min) from classification, SOL_d, for each recording, and the computed SOL time (min) from ground truth (GT) labels, SOL_{GT}, for the same recordings.

BIAS (MIN) - Difference between SOL_d and SOL_{GT}, computed, for each recording, according to equation 3:

$$Bias = SOL_d - SOL_{GT} \tag{3}$$

For a given group of recordings regarding L_1 error or bias values, the arithmetic mean (or just mean), $\overline{x}$, and standard deviation, σ, are computed according to equations 4 and 5, respectively [53]. The arithmetic mean is a measure of central tendency which is used in this work with the intent of summarizing statistics on a certain metric. The standard deviation is computed with the goal of addressing how spread out the data is.

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{5}$$

Where x_i is either a value regarding L_1 error or bias, for a given recording i, in a group of recordings, of size N.

In order to visualize the comparison of SOL-detection performance between different approaches of classification, the **Bland-Altman plot** [54] will be used. It aims at providing an answer on whether two methods are comparable enough to the point where one could replace the other. This kind of plot is more suitable for method comparison than the obvious plot of one method against the other, because with the latter, the

greater the range of the measurements, the better will the agreement seem to be [54]. Thus, it is more appropriate to represent the difference between two measurements, of the same measure, against the mean between those measurements, as shown on figure 11. In this way, it becomes much easier to assess the magnitude of disagreement, regarding both L_1 error and bias, besides allowing an instant identification of outliers and trends [54].

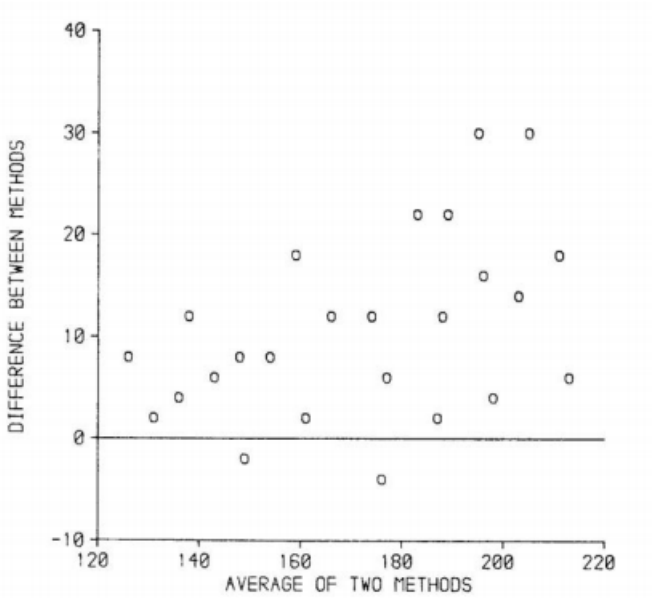


Figure 11.: Sample Bland-Altman plot for method comparison [54].

When in a situation of bias comparison between several methods, for ease of visualization, a box-and-whisker plot [55] is used.

- In the case of the overall epoch by epoch classification performance, all measures are deduced from the **confusion matrix**, which is exemplified on figure 12.

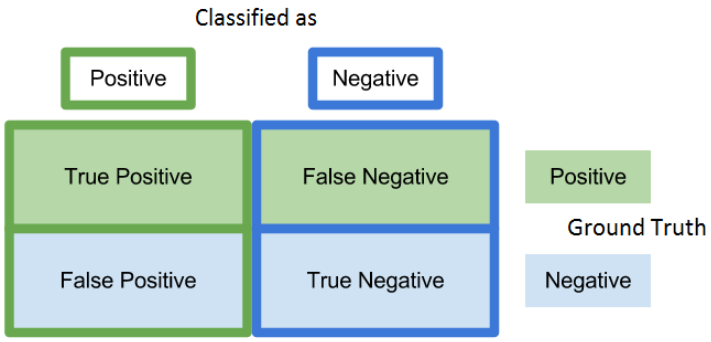


Figure 12.: Example confusion matrix for two possible outcomes: positive and negative.

A true positive (TP) corresponds to the situation when the outcome of the classification is the positive class (defined in this work as *wake*) and it matches the GT label. A true negative (TN) corresponds to the opposite scenario: a negative outcome (*sleep*) matches the GT label. If, however, a positive output happened for a negative GT label, it would be a false positive (FP), also known as a *type I error*. Likewise, a negative classification outcome with a positive GT label correspondence is a false negative (FN), also denominated a *type II error*.
The statistical measures of a binary classification test that are computed from a confusion matrix and to be mentioned in this report are the following:

SENSITIVITY: a measure of the goodness of a test at detecting positive outcomes. It is also called recall and/or true positive rate (TPR). Mathematically, it is the fraction of correctly classified positive instances:

$$Sensitivity = \frac{TP}{P} \tag{6}$$

Where $P = FN + TP$.

SPECIFICITY: a measure of the goodness of a test at detecting negative outcomes. It is also called true negative rate (TNR).

$$Specificity = \frac{TN}{N} \tag{7}$$

Where $N = FP + TN$.

PRECISION: a measure of the reliability of the positive outcomes. It is also called positive predictive value (PPV).

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

ACCURACY: a measure of how often the classification is correct:

$$Accuracy = \frac{TP + TN}{P + N} \tag{9}$$

In the case of evaluation of overall performance, accuracy is one of the most commonly used parameters, because it describes the agreement between 2 sequences of labels. However, it should only be used when both classes are numerically balanced or have the same importance, otherwise, it could be a misleading indicator of performance. **Class imbalance** refers to classification problems where the classes are not represented equally. This is often the case in *sleep/wake* detection. Commonly, over the PSG recordings of a night, one class (*sleep*) is largely represented over the other. Suppose, for the pooled case, that 90% of the epochs belong to *sleep*: a trivial classification output would be that all instances are classified as *sleep*, the majority class. This classification would then have an accuracy of 90%, even though the classification performance over the minority class, *wake*, is unacceptable. A suitable measure for the case of class imbalance is the **Cohen's kappa** (κ) coefficient of agreement [56]. It is an indicator of the level of agreement between 2 sequences of labels when corrected by chance, i.e., it reveals how much better/worse the classification result is than what would be randomly expected. The formula to calculate the κ coefficient of agreement is given by the following equation:

$$\kappa = \frac{obs_{ag} - chance_{ag}}{1 - chance_{ag}} \tag{10}$$

Where obs_{ag} is the proportion of observations in agreement and $chance_{ag}$ is the proportion in agreement due to chance.
By expecting some agreement to randomly occur and subtracting it from the observed agreement, kappa becomes robust to the problem of class imbalance. For this reason, it provides a better understanding of the performance of *sleep/wake* detection. The classifier decision-making threshold was chosen so that it would allow maximum pooled kappa on the training set [57].
Kappa values can vary from -1 to 1. Table 5 offers an interpretation organized in ranges.

Table 5.: Interpretation of Cohen's κ values, according to literature [58].

κ	**Agreement**
< 0	Poor
$0.01 - 0.20$	Slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 0.99$	Almost perfect

3.2 CURRENT PERFORMANCE ASSESSMENT

Following the guidelines described in section 3.1.3, all recordings presented on table 1 were submitted to a binary *sleep/wake* stage classification, over the TIB length of each recording, using an existing *sleep/wake* classifier.

After classification (using a 10-fold cross validation method), the subject pool was logically analyzed in two classes, according to GT information: one constituted by subjects who have normal (good) sleep and the other constituted by troubled sleepers. The criteria applied to recognize such subjects took into consideration SE and SOL time. Troubled sleepers were considered those who had $SE < 85\%$ or/and $SOL_{GT} > 30$ min, while good sleepers were recognized as the remaining subjects, who did not fit the criteria. The results regarding the joint group (complete pool of subjects) are also presented in this analysis. The distribution of subjects into groups was done so that, in a way, the degree of the similarity between two subjects on the same group is high.

General statistics regarding sleep standard measurements for each group were collected and are described on table 6, as well as the performance measurements obtained from classification.

Table 6.: General mean and standard deviation values over sleep statistics on the complete pool, with emphasis on the troubled and good sleepers, as well as current classification performance measurements. For the kappa metric and conventional statistic performance measures the pooled value is given, followed by mean and standard deviation, between brackets.

Group	**Pool (All subjects)**		**Troubled Sleepers**		**Good Sleepers**	
Parameter	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**
N	339 recordings		212 recordings		127 recordings	
TIB (hour)	7.87±0.56	4.23-9.33	7.90±0.53	5.32-9.33	7.81±0.59	4.23-9.21
SOL_{GT} (min)	23.47±21.50	1.50-141.00	30.37±24.18	2.50-141.00	11.96±6.87	1.50-28.50
SE (%)	80.76±11.76	21.39-97.89	74.59±10.62	21.39-91.37	91.05±3.35	85.05-97.89
SOL_d **(min)**	19.73±13.83	0.50-117.50	22.33±16.02	0.50-117.50	15.41±7.28	0.50-56.00
L_1 **error (min)**	10.03±14.94	0.00-104.00	12.55±18.09	0.00-104.00	5.82±4.74	0.00-30.00
Bias (min)	(-3.74)±17.61	(-104.00)-35.50	(-8.04)±20.50	(-104.00)-35.50	3.45±6.68	(-16.00)-30.00
κ	0.61 (0.59 ±0.15)	0.05-0.88	0.63 (0.62±0.15)	0.05-0.88	0.53 (0.54±0.14)	0.14-0.59
Sensitivity	0.68 (0.72 ± 0.17)	0.14-1.00	0.78 (0.68 ± 0.17)	0.16-1.00	0.79 (0.79 ± 0.14)	0.14-1.00
Specificity	0.92 (0.93 ± 0.06)	0.67-1.00	0.89 (0.93 ± 0.05)	0.68-1.00	0.91 (0.91 ± 0.06)	0.67-0.98
Accuracy	0.88 (0.88 ± 0.07)	0.34-0.97	0.87 (0.87 ± 0.07)	0.34-0.97	0.90 (0.90 ± 0.05)	0.70-0.97
Precision	0.68 (0.63 ± 0.20)	0.11-1.00	0.66 (0.78 ± 0.14)	0.27-1.00	0.47 (0.50 ± 0.16)	0.11-0.78
FPR	0.08 (0.07 ± 0.06)	0.00-0.33	0.11 (0.07 ± 0.05)	0.00-0.32	0.09 (0.09 ± 0.06)	0.02-0.33

The L_1 error obtained seems to be much higher for troubled sleepers, in comparison to good sleepers. The bias measurement indicates that, on average, this is an error of underestimation for troubled sleepers (which have a great influence on the general pool results, having, as well, underestimated SOL_d), while it is an error of overestimation for good sleepers. It is quite alarming that the range of estimated bias reaches up to 104 min of underestimation (encountered value for a troubled sleeper). In order to access

these differences in a more visual manner, figure 13 represents the bias estimation on SOL-detection, by means of a Bland-Altman plot.

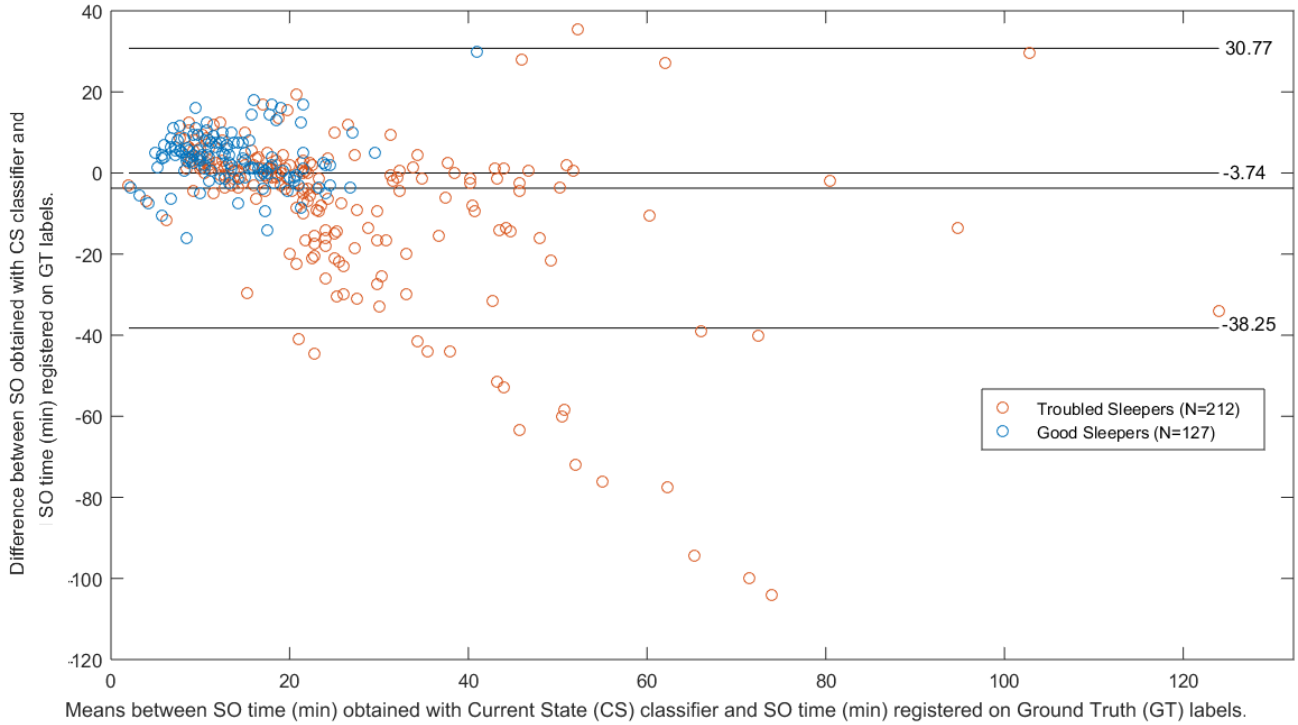


Figure 13.: Bland-Altman plot on SOL-detection error obtained from current classification. Each point represents the estimated bias on SOL-detection for a certain subject (N=339). The 3 horizontal lines represent (from top to bottom) the upper limit of agreement; average bias; and lower limit of agreement.

As it can be seen, the error increases with increasing SOL_{GT}, which highlights the underestimation explained above for troubled sleepers: there is a clear proportional bias. On average there is an underestimation of 3.74 min on SOL-detection.
The upper and lower limits of agreement are, respectively 30.77 min and (-38.25) min. There are several values out of the interval considered by these limits (mainly under the lower limit), which identify the presence of outliers (all belonging to the category of troubled sleepers). It becomes even more obvious that these are the subjects over which an improvement on SOL-detection is more desired, since the results regarding them are having a strongly negative influence over the results on the complete (mixed) pool.
In order to conduct a more detailed analysis, the group of troubled sleepers was further investigated, since classification and SOL-detection performance might differ between people with delayed SOL and people with difficulties in maintaining sleep (low SE). On figure 14 there is a Venn diagram representing the logical sets of subjects that exist within the group of troubled sleepers.

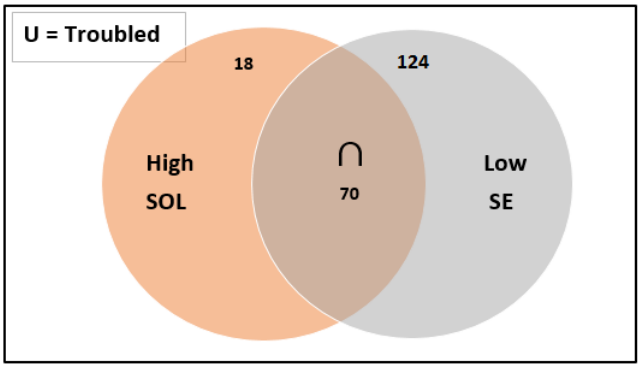


Figure 14.: Venn diagram representing different types of troubled sleepers within the complete universe of troubled sleepers (size 212).

Within troubled sleepers, approximately 33% demonstrated having both low SE and high SOL, while around 58.5% have exclusively low SE issues and only 8.5% have exclusively high SOL. When not considering the distinction regarding to exclusivity, 41.51% of troubled sleepers revealed suffering from high SOL and 91.51% from low SE. Accordingly, the results regarding the following groups of troubled sleepers were detached from the overall results for a closer analysis:

1. Intersection: Recordings with presence of both low SE and high SOL problems (N=70);
2. High SOL: Recordings with presence of high SOL problems (N=88);
3. Exclusively High SOL: Recordings with presence of only high SOL problems (N=18);
4. Low SE: Recordings with presence of low SE problems (N=194);
5. Exclusively low SE: Recordings with presence of only low SE problems (N=124);

General sleep statistics and current classification performance measures on the several kinds of troubled sleepers are presented on tables 7 and 8. The Bland-Altman plot of the bias estimation can be seen on figure 15 [1].

From the analysis of table 7 and figure 15 it becomes even more evident that the current SOL-detection error is higher for subjects who experience delayed SOL_{GT}. In fact, subjects who had exclusively low SE are concentrated closer to the region of $y = 0$ (the average bias is 1.17 min), having a maximum error difference of 19.50 min of SOL overestimation and 11.50 min of underestimation. When adding to these subjects the remaining low SE sleepers (the intersection group) the values on average error rise dramatically from 4.38 min to 12.81

1 Figure 15 is different from figure 13 because it does not include the results regarding the group of good sleepers. It is a representation of the bias on SOL-detection, with current methods, regarding only the group of troubled sleepers. Hence, the average bias value and the values of upper and lower limits of agreement are different from those on figure 13. Also, in the legend, we distinguish between 3 sub groups of troubled sleepers, to address which kind of sleep problem seems to be more associated with higher SOL-detection error.

Table 7.: General sleep statistics on different groups of troubled sleepers, as well as some current classification performance measurements, which are presented by mean and standard deviation values. For the kappa metric the pooled value is given first, followed by the mean and standard deviation values between brackets.

Group	**All troubled sleepers**		**Intersection sleepers**		**High SOL sleepers**	
Parameter	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**
N	212 recordings		70 recordings		88 recordings	
TIB (hour)	7.90±0.53	5.32-9.33	7.92±0.53	5.32-8.93	7.94±0.52	5.32-8.93
SOL (min)	30.37±24.18	2.50-141.00	55.17±25.75	30.00-141.00	51.36±24.33	30.00-141.00
SE (%)	74.59±10.62	21.39-91.37	70.24±12.88	21.39-84.93	73.65±13.33	21.39-91.37
SOL_d **(min)**	22.33±16.02	0.50-117.50	30.99±23.40	0.50-117.50	30.34±21.45	0.50-117.50
L_1 **Error (min)**	12.55±18.09	0.00-104.00	27.78±24.38	0.50-104.00	24.06±23.30	0.00-104.00
Bias (min)	(-8.04)±20.50	(-104.00)-35.50	(-24.19)±27.98	(-104.00)-27.98	(-21.02)±26.11	(-104.00)-35.50
κ	0.63 (0.62±0.15)	0.05-0.88	0.63 (0.62±0.17)	0.07-0.87	0.64 (0.64±0.16)	0.07-0.88

Table 8.: General sleep statistics on different groups of troubled sleepers, as well as some current classification performance measurements, which are presented by mean and standard deviation values. For the kappa metric the pooled value is given first, followed by the mean and standard deviation values between brackets.

Group	**Exc.High SOL sleepers**		**Low SE sleepers**		**Exc.Low SE sleepers**	
Parameter	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**
N	18 recordings		194 recordings		124 recordings	
TIB (hour)	8.02±0.47	7.08-8.83	7.89±0.54	5.32-9.33	7.88±0.54	5.76-9.33
SOL (min)	36.53±6.77	30.50-52.00	25.12±2.50	2.50-141.00	15.47±6.72	2.50-29.50
SE (%)	86.90±1.73	85.09-91.37	73.45±10.37	21.39-84.93	75.25±8.16	45.73-84.79
SOL_d **(min)**	27.83±11.04	11.00-44.50	21.81±16.34	0.50-117.50	16.64±6.04	0.50-36.00
L_1 **Error (min)**	9.70±9.47	0.00-30.00	12.81±18.68	0.00-104.00	4.38±3.89	0.00-19.50
Bias (min)	(-8.69)±10.44	(-30.00)-4.50	(-7.98)±21.21	(-104.00)-35.50	1.17±5.75	(-11.50)-19.50
κ	0.70 (0.70±0.10)	0.49-0.10	0.62 (0.62±0.15)	0.05-0.87	0.62 (0.61±0.15)	0.05-0.86

min and the new maximum error found is of 104 min of SOL-underestimation.

The situation is even more aggravated for the case of high SOL sleepers. Starting with the exclusively high SOL sleepers, the results on average error and bias are worse than those obtained for exclusively low SE subjects. There is an average underestimation of 8.69 min. When considering also the remaining high SOL sleepers (intersection group), the average bias moves to 21.02 min of underestimation, and the absolute error goes from 9.70 min to 24.06 min.

In general, for 212 recordings of troubled sleepers, there is an average SOL-underestimation of 8.04 min, ranging from the extremes of 104 min of underestimation to 35.50 min of overestimation. These results are highly influenced by the errors made for subjects which are not exclusively low SE sleepers, even though more than half of all troubled sleepers belong to the latter (124 recordings out of 212).

Figure 16 represents the percentage distribution of the major distinct groups of subjects considered within the universal set. As demonstrated, only 37% of the recordings belong to good sleepers. The remaining recordings account for delayed SOL or/and low SE.

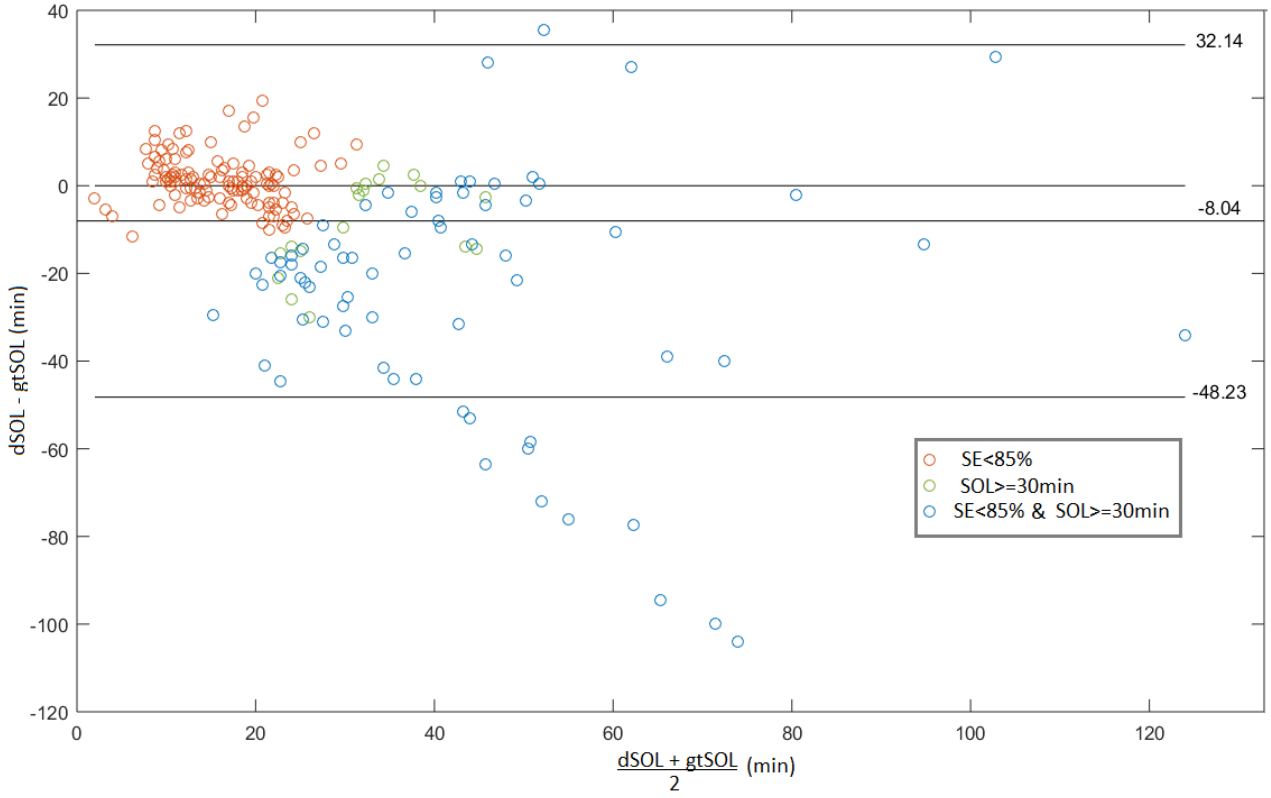


Figure 15.: Bland-Altman plot on SOL-detection error obtained with current classification methods. Each point represents the bias on SOL-detection for each subject on the troubled group (N=212). The 3 horizontal lines represent (from top to bottom) the upper limit of agreement; average bias; and lower limit of agreement.

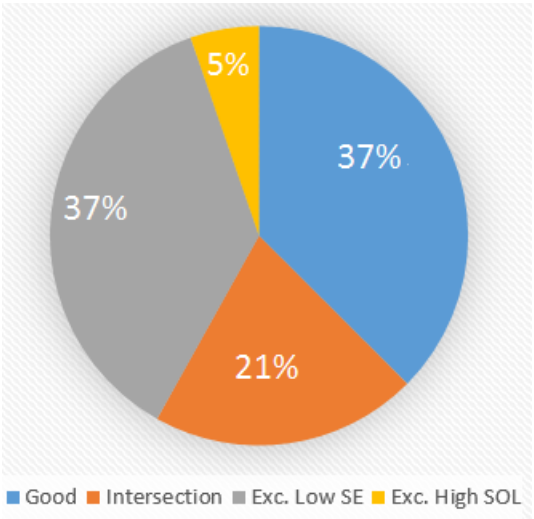


Figure 16.: Pie chart representing subjects' relative distribution within the universal set of size 339.

4

EXPERIMENTAL STUDY ON THE PERFORMANCE OF SOL DETECTION

The goal of this chapter is to perform an experimental study on how certain aspects of the classification process influence the results on SOL detection. Specifically, we will manipulate variables (such as the time of the analysis, usual quality of sleep of the subjects, feature transformation methodologies and FS) to understand how the performance of SOL detection (dependent variable) is affected. Also, we compare the performance obtained with altered approaches for *sleep/wake* detection and with the current classification method presented in section 3.2.

4.1 DATA SET AND METHODS

For the analysis described in the following sections of this chapter, we used the data set presented in section 3.1.1 for a *sleep/wake* scoring, performed according to the framework described in section 3.1.3. The data set was partitioned in groups as suggested on table 6 (section 3.2) before the classification process.
The following independent variables are manipulated:

1. Time of the analysis: contrary to current classification methods presented in section 3.2, where an epoch-by-epoch classification was performed on each recording since the moment of lights off until the moment of lights on, the present study, identifies the ending time point of the classification as the moment corresponding to: SOL_{GT} plus 30 additional epochs (15 minutes);

2. Training/testing population: Before classification, the pool of subjects was sorted in groups. It should be noted that recordings of the same subject were carefully kept in the same group. Then, a 10-fold-cross validation was performed separately for each one of the considered groups. This is different from the separation of subjects presented in section 3.2, even though the criteria utilized was the same. In section 3.2 classification was performed one time for the entire pool. Only then the results were analyzed separately between groups of subjects;

3. Feature transformation: as a first approach the classification was performed using only feature normalization techniques for feature transformation, while only in a second approach both types of transformation methods (normalizing and post-processing) were applied. Measures of performance regarding SOL-detection and *sleep/wake* detection, for both scenarios of feature transformation are compared, within the same group of subjects, with the goal of determining if it is advantageous to use post-processing methods, considering its usage has the disadvantage of over-transforming the data;

4. Feature restriction (FR): In this study *sleep/wake* scoring was repeated using three different sets of features, as feature subsets for the classification step. Firstly, the resulting set from a CFS FS, as presented in section 3.1.3, was used. Then, in order to gather characteristics of interest for SOL-detection purposes without having to proceed with the partition of the data set in more groups than into good and troubled, both intersection and union of the selected features (regarding all troubled sub-groups mention in the previous section 3.2, and presented on figure 14) were performed. The intersection set resulted in 13 features, for normalizing-only case, and 24 features when both transformations were applied. The union set comprised a total of 29 features (normalizing-only case) and 62 features (both transformations applied-case).
   After collecting the feature sets, classification was performed using these sets of features for FR. The performance regarding SOL-detection was then compared between the three approaches and also to the performance of current classifiers.

4.2 RESULTS

4.2.1 *Feature Transformation*

The results on the performance of the classification obtained for each group of subjects are presented on table 9. Within the same group of subjects the performance obtained when varying methods of feature transformation is compared in terms of significance. It is important to notice that the kappa values presented are only referent to the considered time span, and are stated here, only for the purpose of comparing the both approaches of feature transformation for the task of differentiating *sleep* from *wake*. For this reason, the values are not valid to be compared with the ones presented earlier for current classification.

The differences regarding mean and standard deviation values from SOL-detection can be directly visualized on the error bar graph on figure 17. The SOL-detection error comparison for the additional groups referenced in section 3.2 is also included on the same figure.

Table 9.: Between brackets are the arithmetic mean and standard deviation values obtained on κ metric, preceded by the pooled κ value. Also, means and standard deviations values obtained for SOL-detection performance measures are presented, and compared among groups of subjects, and methods of feature transformation (Norm indicates normalization and PP indicates post-processing). Significance of the difference between the results regarding L_1 error and κ coefficient, obtained with different processing methods, for the same group of subjects, was calculated with a two-tailed Wilcoxon-signed rank test [59]. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Group	**Pool**		**Troubled Sleepers**		**Good Sleepers**	
N	339		212		127	
Transformation	**Norm**	**Norm+PP**	**Norm**	**Norm+PP**	**Norm**	**Norm+PP**
κ	0.66 (0.60±0.24)	0.75 (0.65 ± 0.28)***	0.62 (0.57±0.27)	0.71(0.61±0.29)**	0.70 (0.67±0.20)	0.75 (0.72±0.20)**
SOL_d **(min)**	22.47±14.21	24.03±23.31	30.02±17.03	31.21±26.05	12.68±5.50	11.94±5.96
L_1 **error (min)**	7.55±10.20	4.62±5.21***	8.19±10.48	5.49±6.23***	4.19±3.65	3.05±3.12 ***
Bias (min)	(-1.0)±12.66	0.56±6.94	(-0.35)±13.30	0.85±8.27	0.72±5.52	(-0.01)±4.37

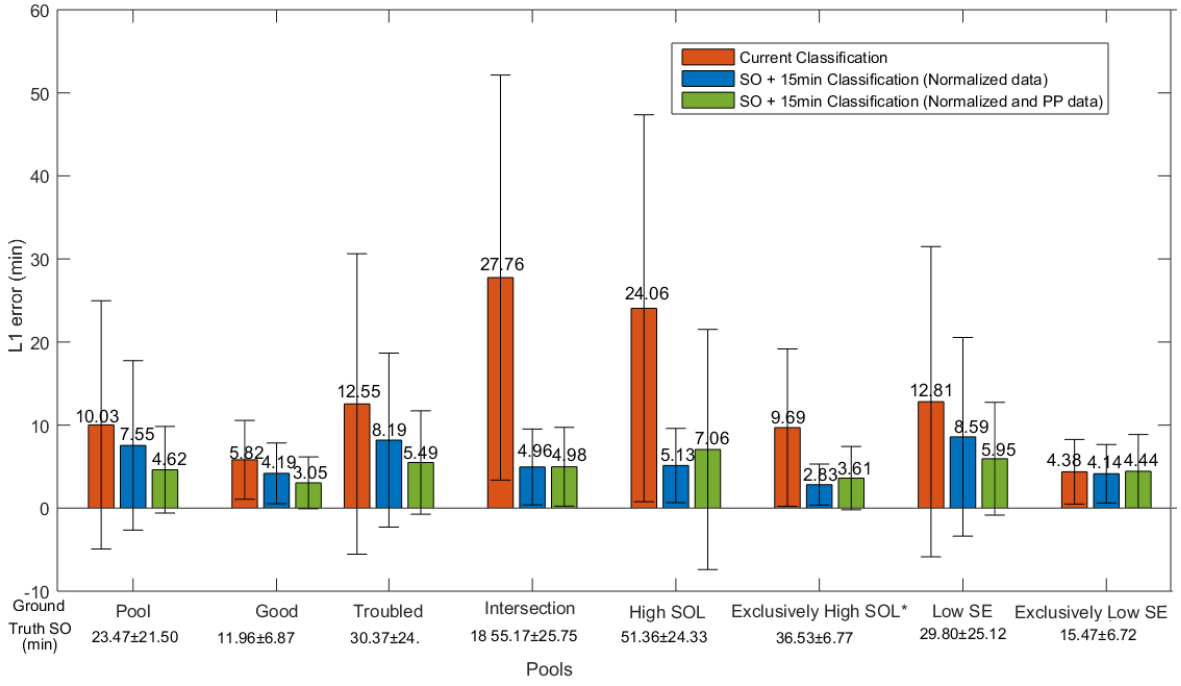


Figure 17.: Mean absolute SOL-detection error and standard deviations obtained for each group. Under the x-axis label are represented the values regarding mean and standard deviation of SOL_{GT} time, regarding each group. * Given the reduced size of the sample on the exclusively high SOL group, the classifier used to test this group was trained on the intersection group.

4.2.2 *Feature Restriction*

Figures 18 and 19 illustrate the results on mean L_1 errors (min) of SOL-detection obtained for the main groups of subjects, when altering the feature sets between the one resulting from CFS FS (as presented in section 3.1.3); and the intersection and union sets for FR. Also,

the values obtained are compared between methods of feature transformation and to the current classification.

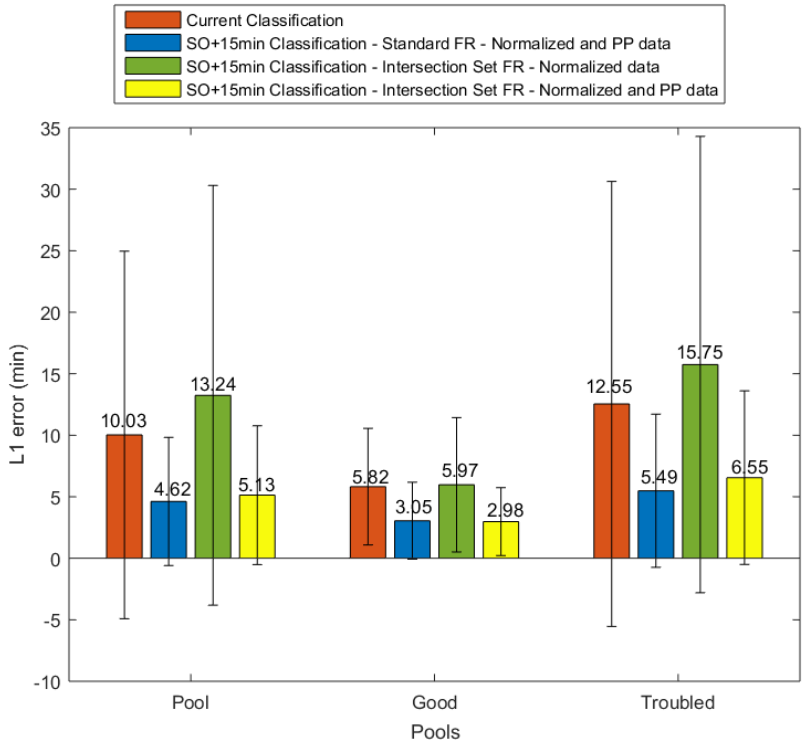


Figure 18.: Mean and standard deviation values for the SOL-detection L_1 error obtained for each group, for comparison between current and new methods of classification, using features restricted to the intersection set.

Table 10 summarizes the differences between the 4 methods regarding mean, standard deviation and range values for SOL_d, L_1 error and estimated bias, for the complete pool (N=339). Only the results regarding the application of both methods of feature transformation are presented. Tables 11 and 12 include the results of the same analysis for groups of troubled and good sleepers, respectively.

Table 10.: Mean and standard deviation values obtained for SOL-detection performance over 339 recordings of mixed subjects, when varying the approach of *sleep/wake* classification. Significance of the difference between the results regarding L_1 error obtained with current classification and remaining methods calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Method	**Current**		**FR:Standard FS**		**FR:Intersection set**		**FR:Union set**	
Parameter	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**
Transformation	Normalization + Post-Processing							
SOL_d **(min)**	19.73±13.83	$0.50 - 117.50$	24.03 ± 23.31	$0.50 - 156.00$	23.43 ± 23.25	$3.00 - 156.00$	23.25 ± 23.19	$0.50 - 141.00$
L_1 **Error (min)**	10.03±14.94	$0.00 - 104.00$	$4.62 \pm 5.21^{***}$	$0.00 - 41.00$	$5.13 \pm 5.64^{***}$	$0.00 - 36.50$	$4.38 \pm 4.82^{***}$	$0.00 - 32.50$
Bias (min)	(-3.74)±17.61	$(-104.00) - 35.50$	0.56 ± 6.94	$(-41.00) - 18.00$	$(-0.04) \pm 7.63$	$(-36.50 - 18.00)$	$(-0.22) \pm 6.51$	$(-32.50) - 17.00$

Figures 20, 21 and 22 illustrate a closer comparative analysis of the bias estimation on SOL-detection obtained by means of a box-and-whisker plot, respectively for the complete

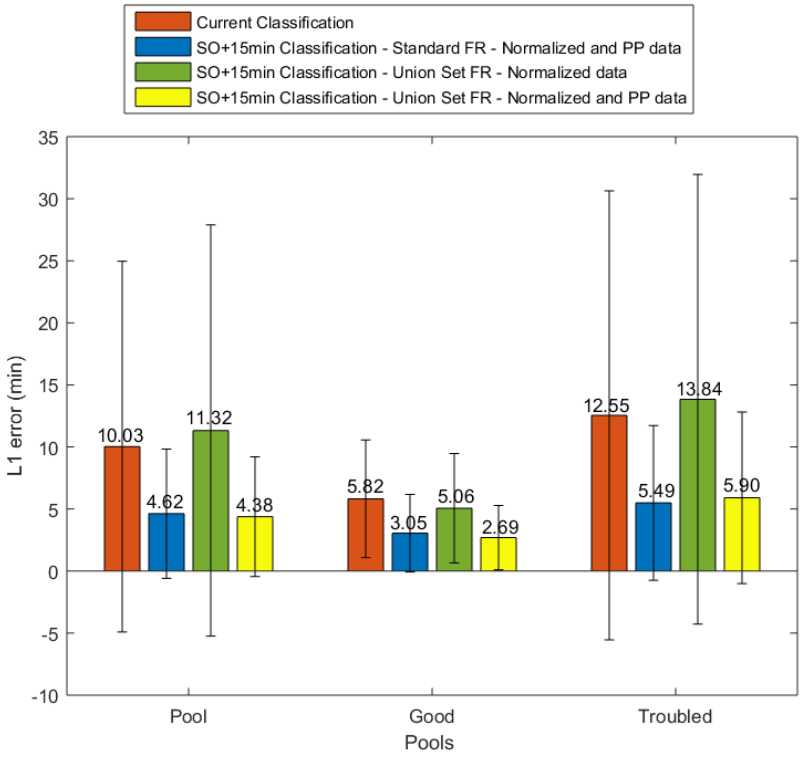


Figure 19.: Mean and standard deviation values for the SOL-detection L_1 error obtained for each group, for comparison between current and new methods of classification, using features restricted to the union set.

Table 11.: Mean and standard deviation values obtained for SOL-detection performance over 212 recordings of troubled sleepers, when varying the approach of *sleep/wake* classification. Significance of the difference between the results obtained with current classification and remaining methods, regarding L_1 error, was calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Method	**Current**		**FR:Standard FS**		**FR:Intersection set**		**FR:Union set**	
Parameter	$\overline{x} \pm \sigma$	**Range**	$\overline{x} \pm \sigma$	**Range**	$\overline{x} \pm \sigma$	**Range**	$\overline{x} \pm \sigma$	**Range**
Transformation	Normalization + Post-Processing							
SOL_d **(min)**	22.33 ± 16.02	$0.50 - 117.50$	31.21 ± 26.05	$5.50 - 156.00$	29.01 ± 26.43	$0.50 - 156.00$	29.91 ± 25.25	$0.50 - 144.50$
L_1 **Error (min)**	5.49 ± 6.23	$0.00 - 104.00$	$5.49 \pm 6.23^{***}$	$0.00 - 61.50$	$6.55 \pm 7.05^{***}$	$0.00 - 36.00$	$5.90 \pm 6.91^{***}$	$0.00 - 42.00$
Bias (min)	$(-8.04) \pm 20.50$	$(-104.00) - 35.50$	0.85 ± 8.27	$(-61.50) - 18.50$	$(-1.35) \pm 9.54$	$(-36.00) - 15.00$	$(-0.46) \pm 9.09$	$(-42.00) - 18.50$

Table 12.: Mean and standard deviation values obtained for SOL-detection performance over 127 recordings of good sleepers, when varying the approach of *sleep/wake* classification. Significance of the difference between the results obtained with current classification and remaining methods, regarding L_1 error, was calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Method	**Current**		**FR:Standard FS**		**FR:Intersection set**		**FR:Union set**	
Parameter	$\overline{x} \pm \sigma$	**Range**	$\overline{x} \pm \sigma$	**Range**	$\overline{x} \pm \sigma$	**Range**	$\overline{x} \pm \sigma$	**Range**
Transformation	Normalization + Post-Processing							
SOL_d **(min)**	15.41 ± 7.28	$0.50 - 56.00$	11.94 ± 5.96	$3.00 - 28.50$	12.13 ± 5.75	$1.50 - 29.50$	11.42 ± 6.23	$1.50 - 32.50$
L_1 **Error (min)**	5.82 ± 4.74	$0.00 - 30.00$	$3.05 \pm 3.12^{***}$	$0.00 - 17.00$	$2.98 \pm 2.76^{***}$	$0.00 - 17.00$	$2.69 \pm 2.59^{***}$	$0.00 - 11.00$
Bias (min)	3.45 ± 6.68	$(-16.00) - 30.00$	$(-0.01) \pm 4.37$	$(-17.00) - 15.50$	0.17 ± 4.07	$(-10.50) - 17.00$	$(-0.54) \pm 3.70$	$(-11.00) - 7.00$

pool of subjects (N=339), for the troubled group (N=212) and for the group of good sleepers (N=127). Here are stated the values regarding only the case of both feature transformation methods.

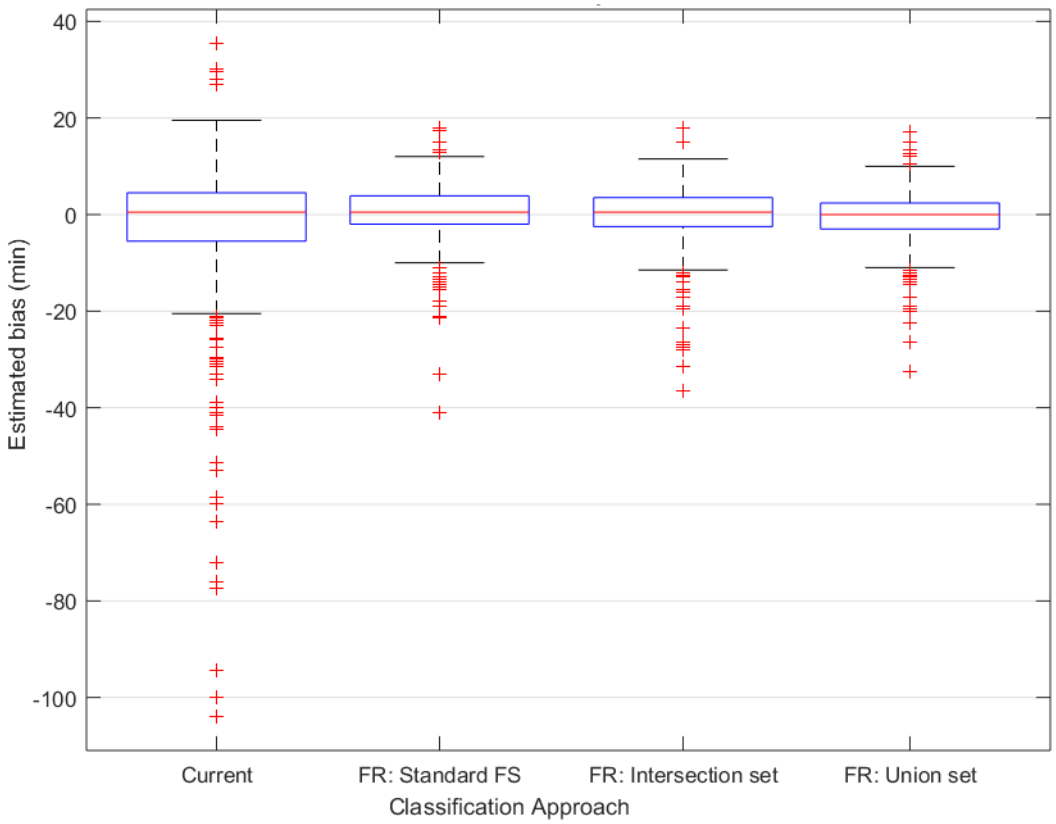


Figure 20.: Bias obtained for the pool of subjects (N=339) when varying the *sleep/wake* classification approach between: currently in use; considering only the beginning of the recordings and a)performing CFS FS; b)restricting features for training/testing to the intersection set; c) restricting features for training/testing to the union set.

On figure 23 there is a representation on the estimation of bias obtained when restricting the time of the analysis and simultaneously applying the union set of features for FR, by means of a Bland-Altman plot.

These results address the SOL-detection error as being any value of obtained SOL time that does not correspond exactly to the SOL_{GT} time, for a given subject, which might be a too strict definition for *error* in the clinical context of the problem. This observation motivated the study of how the error percentage obtained for each approach is affected when varying the definition of error. The variation was done for values between zero and 15 minutes (incremented by 1 minute) as the time difference between SOL_d and SOL_{GT} above which the existence of an L_1 error is indeed acknowledged. Figure 24 demonstrates the results of the analysis, for the complete pool of subjects.

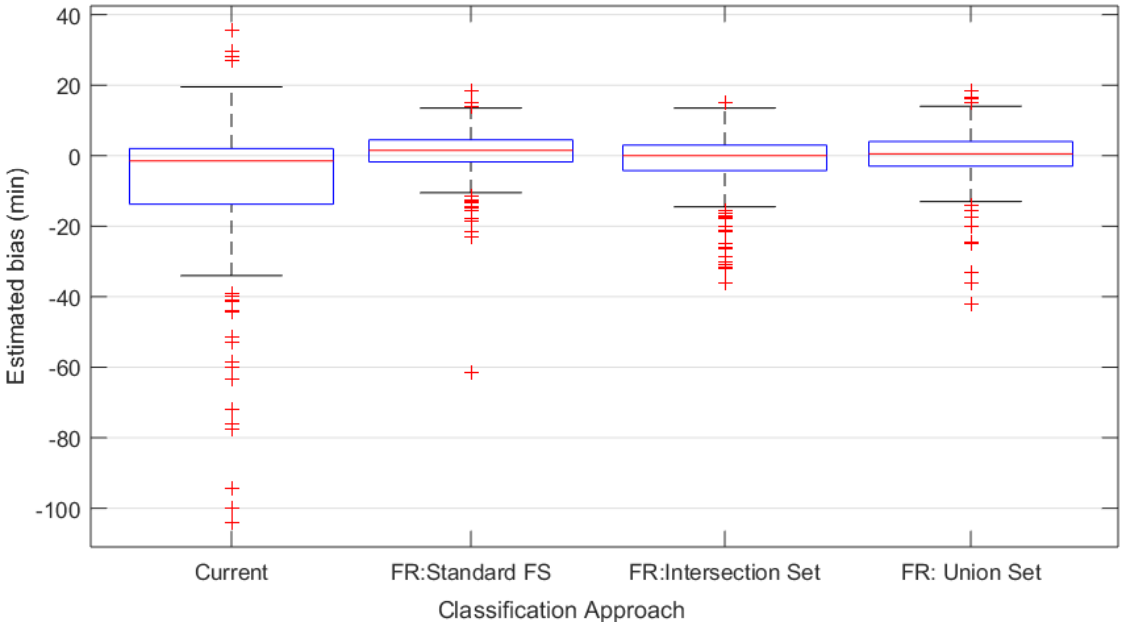


Figure 21.: Bias obtained for the group of troubled sleepers (N=212) when varying the *sleep/wake* classification approach between: currently in use; considering only the beginning of the recordings and a)performing standard FS; b)restricting features for training/testing to the intersection set; c) restricting features for training/testing to the union set.

4.3 DISCUSSION

Table 9 shows that, for this time-restricted analysis, it is beneficial to apply post-processing methods (on top of normalization methods) in the step of feature transformation. Within each group of subjects considered, the increase on the Cohen's kappa value is notorious and significant (after a two-tailed Wilcoxon signed-rank test), for all of them. On average, there was an increase of 8.00%, 6.70% and 6.72% for the original, troubled and good sleepers groups, respectively. Also the pooled value has increased above 0.70 for all groups, indicating substantial agreement.

Regarding the results on SOL-detection performance, there were also major improvements with the application of post-processing techniques (these also proved to be significant after a two-tailed Wilcoxon-signed rank test). Specifically, the average L_1 error has decreased by 39.23%; 33.61% and 29.51% for the complete; troubled and good pools, respectively. Figure 17 makes the improvements on the reduction of the L_1 errors even more evident. Also, it includes the values from the current classification presented in section 3.2 in the comparison, as well as the results on the sub-groups of troubled subjects. It can be seen that, for almost every group considered (with the exception of the exclusively low SE group), the largest average L_1 error happens for the current classifier, which places it as the less desirable method of sleep scoring approach to use, for the purpose SOL-detection.

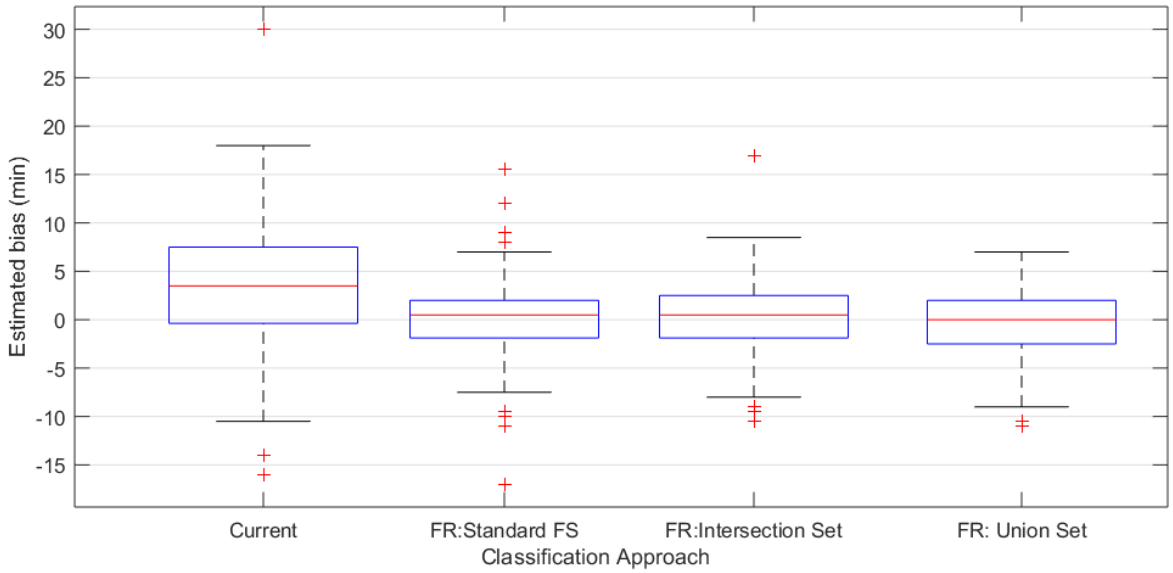


Figure 22.: Bias obtained for the group of good sleepers (N=127) when varying the *sleep/wake* classification approach between: currently in use; considering only the beginning of the recordings and a)performing CFS FS; b)restricting features for training/testing to the intersection set; c) restricting features for training/testing to the union set.

As mentioned in section 4.1, the feature sets resulting from individual classification on the groups of subjects presented on figure 17, were used to create an *intersection* and *union* set of features, which were applied, afterwards, for the task of *sleep/wake* detection on the 3 main groups. On this matter, figures 18 and 19 support that the lowest average L_1 errors on SOL-detection are achieved by using the union set of features for FR (resultant from the usage of both methods of feature transformation). When comparing with the results obtained with current classification methods, which are discriminated on tables 10, 11 and 12, the diminishing of the magnitude of the L_1 errors proved to be significant, for the 3 groups, after a two-tailed Wilcoxon test, at a p-value of 0.001.

There are also improvements on average bias (and standard deviation associated to it), when comparing current classification to any of the new classification approaches. This becomes more clear from the analysis of figures 20, 21 and 22, where it can be visualized that current methods present always the widest range of bias estimation, having outliers over 100 minutes of underestimation, for the complete pool and troubled group. The smallest range and interquartile range (IQR) is obtained on the case of the union set of features for FR, for the complete and good group. For the troubled group, however, a smaller range is obtained with the intersection set of features, even though the average bias is the closest to zero when applying the union feature set.

In general, with new approaches, not only is the range decreased in a large scale, but also the IQR decreases. This is an indicator of diminished variability of SOL-detection L_1 error, with less serious situations of outliers. Still, it is clear the presence of these abnormal cases,

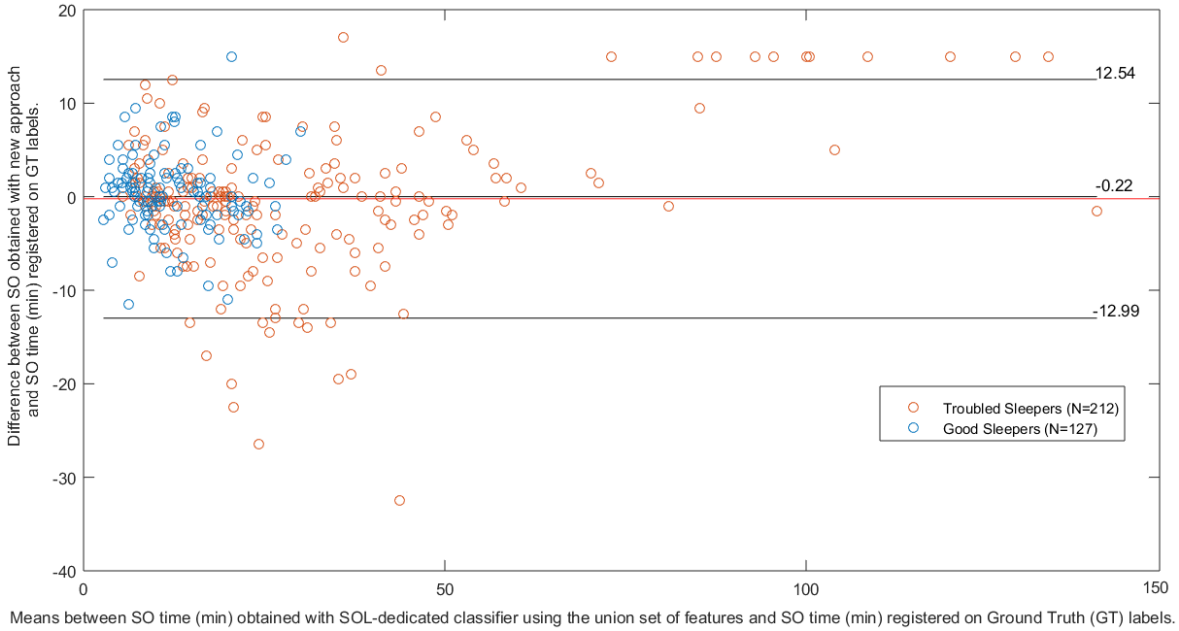


Figure 23.: Bland-Altman on relative SOL-detection error obtained when using the classification method dedicated to the beginning of the recordings combined with the union set of features for training/testing.

which are more frequent on the negative side (over the lower whisker), revealing a general underestimation of SOL.
Regarding the median SOL-detection error value, with the proposed methods of classification, it becomes very close to zero minutes, for every group considered.

Figure 23 underlines even more the benefits in the application of the time-restricted classification, with the usage of the union set, for SOL-detection purposes, in comparison to the currently in use method of classification, for which the bias of SOL-detection was graphically plotted in the same manner, for the same group of subjects (339 mixed recordings) on figure 13, in chapter 3.2.
Obvious distinctions from the analysis of the two figures include the range in limits of agreement that has been reduced from approximately 69 min to approximately 26 min, indicating that the presence of outliers has been greatly reduced, and the errors on SOL-detection are much more closely distributed around the average bias which now indicates an underestimation of only 0.22 min (in comparison to an average of 3.74 min with current methods). Also, with the latter, there was a clear proportional bias that is now inexistent.
One resemblance between both methods is that, regardless of the improvements from one to the other, SOL-detection mistakes of greater magnitude are associated to the classification of troubled sleepers.
Figure 24 shows that, even when varying the definition of error, to make it less strict, one factor remains constant over the increased tolerance in the definition: currently used

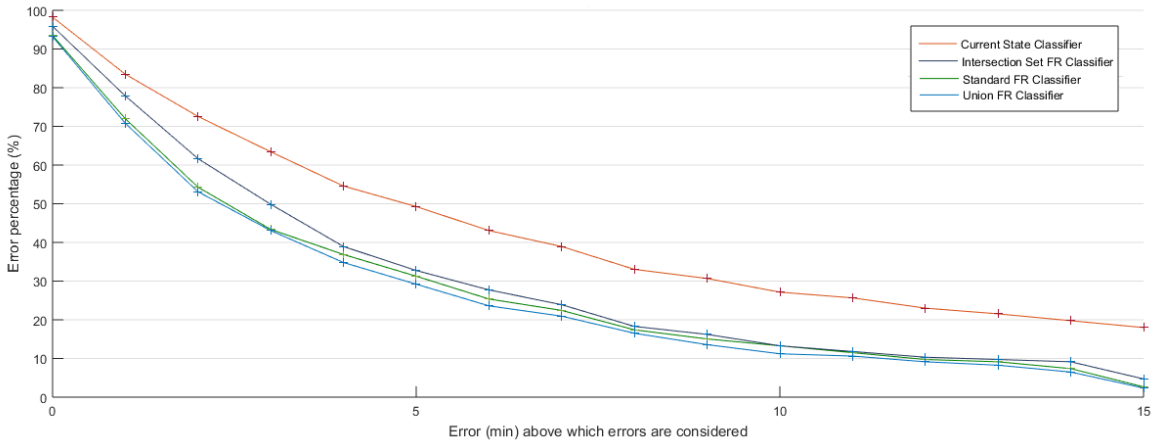


Figure 24.: Error percentage obtained when using 4 different approaches for classification of 339 recordings, according the definition of error considered.

methods of classification provide the highest L_1 error percentage. Besides, it is once again confirmed that *sleep/wake* detection restricted to the beginning of the recordings combined with the usage of a specific set of features (union set) allows to achieve the lowest L_1 error percentage rates, regardless of the level of tolerance on the definition of *error*.
Still, the results obtained with the approach of CFS FS and with the usage of the intersection set of features also outperform current classifiers on the task of SOL-detection, and are close to the ones obtained with the usage of the union set.

5

IMPACT OF ACTIGRAPHY ON THE PERFORMANCE OF SOL DETECTION

Regarding actigraphy-based *sleep/wake* classifiers described in literature, it has been proved that the longer it takes for a subject to fall asleep, the worst the performance of the classification, in particular in the detection of SOL [60]. The explanation for that fact is that, when faced with delayed SOL, a person will tend to behave according to a pattern which mainly consists on lying as still as possible and controlling (calming and regulating) respiration.
Since, as described in section 3.1.3, the classification framework used in this research relies on cardiorespiratory and estimated actigraphy features, most likely, it will often be tricked into overestimating *sleep* and underestimating SOL.
This chapter investigates the influence of the actigraphy feature and correlated features on the performance of SOL-detection.

5.1 DATA SET AND METHODS

For this analysis, the subject pool [1], of size 339, was randomly separated in two sets of approximate size: one for training and the other for testing the classification. It should be noted that in the split of the data, recordings of the same subject, were carefully kept in the same set [2]. Table 13 includes the demographic information on the subjects on both sets.

Recordings from both sets were clipped in different time ranges, from the moment of lights-off until the following 31; 40; 50; 60; 70; 100; 110 and 120 min, with the goal of selecting the most appropriate time window, for increased performance on SOL-detection.
Within the training set, information regarding SOL_{GT} was accessed to further divide the subjects in good and troubled sleepers. Using all the recordings from this two sub-groups, and also from the complete training set, several classifiers were trained. Each classifier was then tested, with examples from the testing set, according to the hold-out method described

1 Demographic information presented in chapter 3, subsection 3.1
2 Recall that most of the subjects on the SIESTA regular data set had their sleep recorded for 2 consecutive nights, as mentioned in section 3.1.1.

Table 13.: Demographic and sleep measurement information regarding subjects on the training and testing sets.

Set	**Training**		**Testing**	
Parameter	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**
N	169 recordings (90 subjects)		170 recordings (90 subjects)	
Sex	45 females (50.0%)		48 females (53.3%)	
Age (year)	49.69 ± 19.24	$20.00 - 86.00$	50.51 ± 20.21	$22.00 - 95.00$
BMI (kg/m^2)	24.65 ± 3.64	$17.16 - 34.84$	24.45 ± 3.36	$16.98 - 35.25$
TIB (hour)	7.89 ± 0.52	$4.23 - 9.33$	7.84 ± 0.59	$5.32 - 9.23$
SOL (min)	22.49 ± 19.69	$1.50 - 141.00$	24.47 ± 23.17	$1.50 - 126.00$
SE (%)	81.17 ± 11.54	$36.95 - 97.89$	80.18 ± 12.34	$21.39 - 97.12$

in section 3.1.3.
Figure 25 illustrates the procedure.

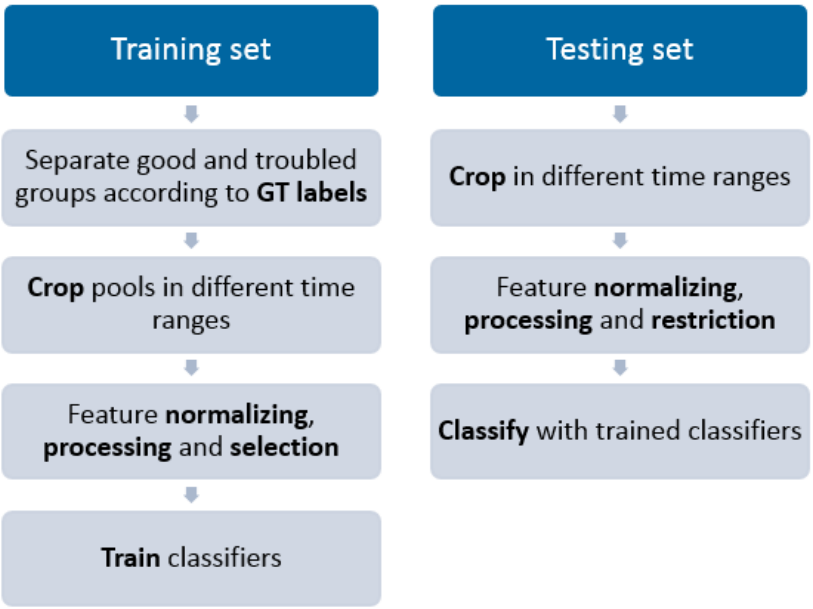


Figure 25.: Model applied for training and testing classifiers for the task of *sleep/wake* detection.

Also, for each classification model, three different feature sets, for FR, were utilized. In total, 72 classification models were trained and tested. Given the dimension of this analysis, the results regarding SOL-detection performance and Cohen's kappa of each model, for each time window considered, are presented in appendix A, for the testing set.
The classification model using CFS FS techniques (described in section 3.1.3) and a time window of 120 min, trained with all the recording examples of the training set, was chosen to be the most adequate to proceed with forward in the analysis. Its usage allowed to obtain the best results on the performance of SOL-detection, as commented in section A.2 of appendix A.

Additionally, three other models were created, in order to access the influence of actigraphy on the performance of SOL-detection. In total, 4 variations (regarding the feature set utilized for FR) of the same model were evaluated:

- M - Classifier trained with examples from the first 120 min of 169 night recordings (training set) and FS performed with a CFS FS algorithm;
- M_0 - Same classification model as in M, except for the absence of the actigraphy feature from the feature set resultant of CFS FS. Table 14 contains the name of each feature on the training set;
- M_1 - Same classification model as in M, except for the absence of the actigraphy feature and strongly correlated features, for which the following condition $|r|>0.6$(where r is the correlation coefficient, was verified with a Pearson's correlation test [61]);
- M_2 - Same classification model as in M, except for the absence the actigraphy feature, strongly and moderately correlated features, for which the following condition $|r|>0.4$ was verified with a Pearson's correlation test [61].

Table 14.: List of features resultant from CFS FS on the training set, for model M. It also contains the indication to the features which were included to train the other models. E - Excluded. I - Included.

Name	M_0	M_1	M_2
Normalized (total) power in the High Frequency band of the respiratory effort signal [5]	I	I	E
Log-std of the respiratory frequency over a sliding window (size 9) [11]	I	I	E
Scaling exponent of detrended fluctuation analysis for slower time scales [62]	I	I	I
Constrained Dynamic Time Warp distance measure (z-score) [63]	I	E	E
Continuous wavelet transform over the respiratory signal. Preserves peaks by subtracting the medfilted signal (z-score, medfilt) [37]	E	E	E
10^{th} percentile of the detrended heart rate	I	I	I
Higuchi measure of phase coordination between respiration and cardiac activity	I	I	I
Number of nodes of interbeat intervals in visibility graph with a small degree (≤ 3) [36]	I	I	I
Mean degree of nodes of breath-to-breathintervals in difference visibility graph [36]	I	I	I
Slope of degree of nodes of cardiorespiratory interaction series in difference visibility graph [36]	I	I	E

Then, the performance of SOL-detection measurements was calculated. The reasoning is summarized by a flowchart diagram, on figure 26.

After performing SOL detection, the performance was compared between models and to the results obtained with current classification methods (presented in section 3.2).

5.2 RESULTS

Table 15 contains the SOL-detection performance results obtained with current methods and with the suggested approaches. The information is organized in groups of subjects, according to SOL_{GT}.

Figures 27, 28 and 29 allow a more visual comparison on bias estimation between classifiers,

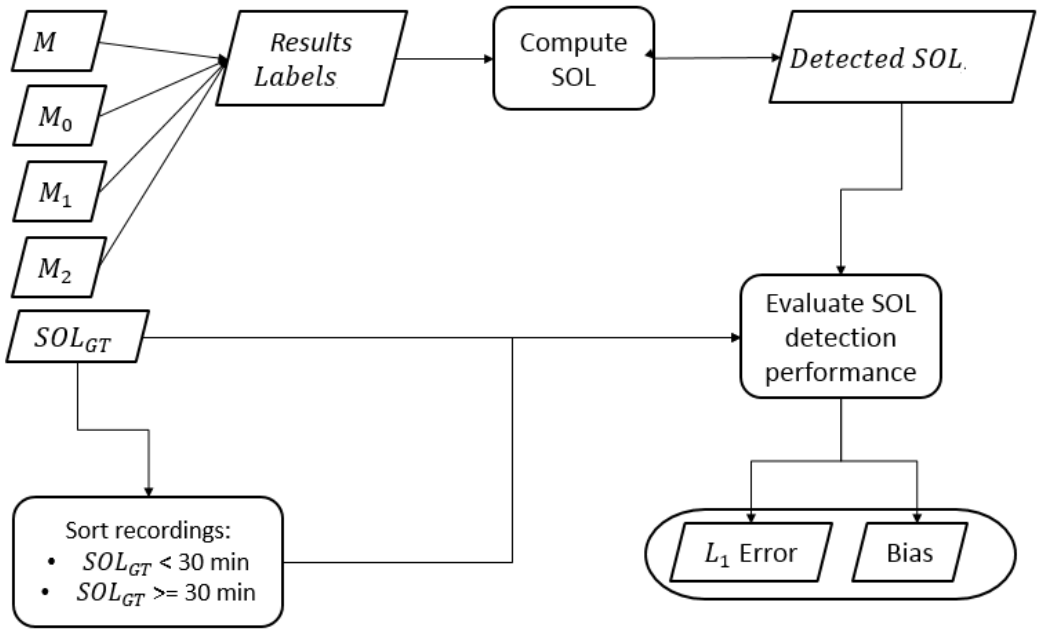


Figure 26.: Flowchart diagram expressing step-by-step of operations performed in this study after obtaining models M, M_0, M_1 and M_2.

for the complete testing set and for subjects with $SOL_{GT} < 30$ min and $SOL_{GT} \geq 30$ min, respectively.

Table 15.: Mean and standard deviation values obtained on the performance of SOL-detection (L_1 error and estimated bias), with different classification models, on the testing set comprised of 170 recordings. Among them, are distinguished those who fit the criteria $SOL_{GT} \in [0, 30[$ min and $SOL_{GT} \in [30, max]$ min. The significance of the differences between the results obtained with new approaches and current classification, regarding L_1 error, within the same group of subjects, was calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Criterion	None		$SOL_{GT} \in [0, 30[$ (min)		$SOL_{GT} \in [30, max[$ min	
N	170		122		48	
$SOL_{GT}(min)$	24.47 ± 23.17		13.41 ± 6.82		52.58 ± 26.18	
Metric	**L_1 error (min)**	**Bias (min)**	**L_1 error (min)**	**Bias (min)**	**L_1 error (min)**	**Bias (min)**
Current	11.49 ± 17.78	$(-5.85) \pm 20.36$	5.01 ± 4.32	2.22 ± 6.25	27.27 ± 26.76	$(-25.58) \pm 28.41$
M	$12.52 \pm 15.64^{**}$	0.79 ± 20.04	$8.78 \pm 8.19^{***}$	7.84 ± 9.10	$22.04 \pm 24.04^{***}$	$(-17.15) \pm 27.82$
M_0	$12.30 \pm 15.14^{**}$	1.05 ± 19.50	$8.76 \pm 9.02^{***}$	7.89 ± 8.89	$21.29 \pm 23.31^{***}$	$(-16.33) \pm 27.09$
M_1	$13.14 \pm 16.60^{***}$	3.21 ± 20.92	$10.11 \pm 11.07^{***}$	9.38 ± 11.70	$20.82 \pm 24.25^{***}$	$(-12.47) \pm 29.53$
M_2	$13.34 \pm 14.95^{**}$	1.49 ± 20.00	$10.70 \pm 8.62^{***}$	9.38 ± 10.05	$20.04 \pm 23.43^{**}$	$(-18.56) \pm 24.63$

5.3 DISCUSSION

From the analysis of table 15 and figure and 27, it becomes clear that, for subjects who take less than 30 min to fall asleep there is no advantage, regarding SOL-detection performance, in replacing the current classification method with any of the remaining methods proposed,

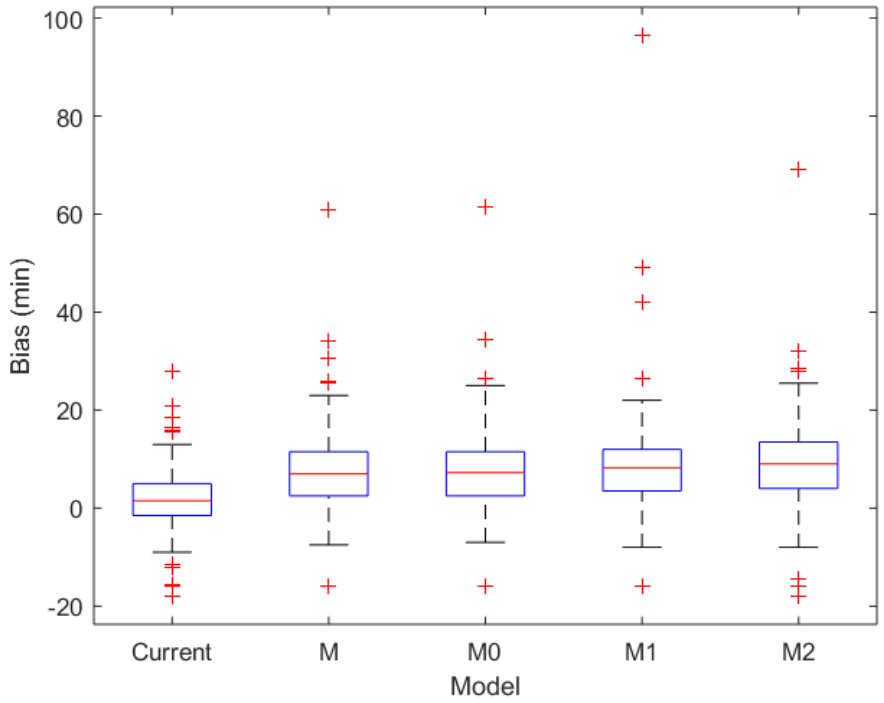


Figure 27.: Box-and-whisker plot on the comparison of bias estimation of SOL with 5 different models of classification, on subjects with $SOL_{GT} < 30$ min (N=122).

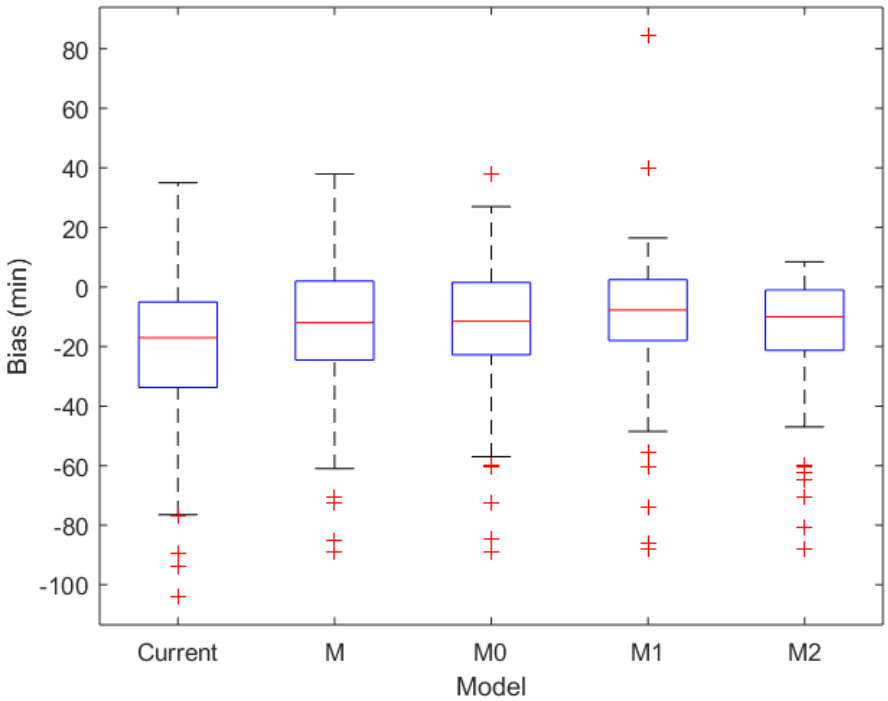


Figure 28.: Box-and-whisker plot on the comparison of bias estimation of SOL with 5 different models of classification, on subjects with $SOL_{GT} \geq 30$ min (N=48).

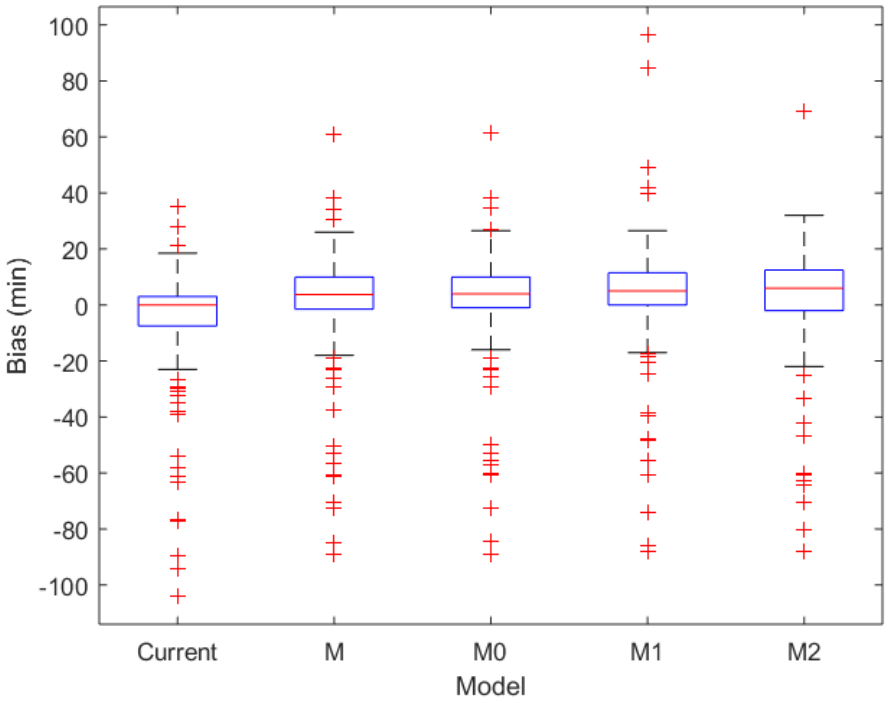


Figure 29.: Box-and-whisker plot on the comparison of bias estimation of SOL with 5 different models of classification, on subjects from the testing set (N=170).

since the average bias obtained with the former is the closest to zero minutes, besides allowing the smallest mean L_1 error, and also the smallest standard deviation interval on this metric. The IQR and range are also the smallest achieved, in comparison to the remaining approaches. Also, the values obtained with new approaches proved to be significantly worse than the one obtained with current classifiers, after a two-tailed Wilcoxon-signed rank test.

However, that is not the case for subjects with $SOL_{GT} \geq 30$ min. For this group, from the values presented on table 15, it seems to be more appropriate to perform SOL detection with the classification model M_2. Comparing to current classifiers, the average value on L_1 error has dropped from 27.27 min to 20.04 min (significant difference with a p-value of 0.01). Even though general underestimation of SOL still persists, it has diminished from an average (-25.58) min to (-18.56) min, accompanied by improvements on the standard deviation of the bias as well, which is related to the elimination of the most extreme outlier values. This is more easily visualized on the box-and-whisker plot on figure 28.

Figure 29 shows that, when trying to analyze the effect of each approach on the complete testing set, it becomes harder to distinguish the most advantageous one. Thus, supporting the idea of dividing the subjects according to SOL_{GT}, as the most appropriate method of classification seems to vary as a function of that variable.

6

PROPOSED REASONING FOR SLEEP/WAKE DETECTION

Based on the insights gained from previous results and discussions, on sections 5.2 and 5.3, this chapter suggests guidelines on the implementation of a new reasoning for *sleep/wake* and SOL detection, which takes into consideration the time of SOL_{GT}.
The results on the performance of SOL-detection and overall classification performance are evaluated, and compared to those obtained with current methods (presented in 3.2).

6.1 DATA SET AND METHODS

Figure 30 exemplifies a real-world situation, for *sleep/wake* detection, with the proposed method.

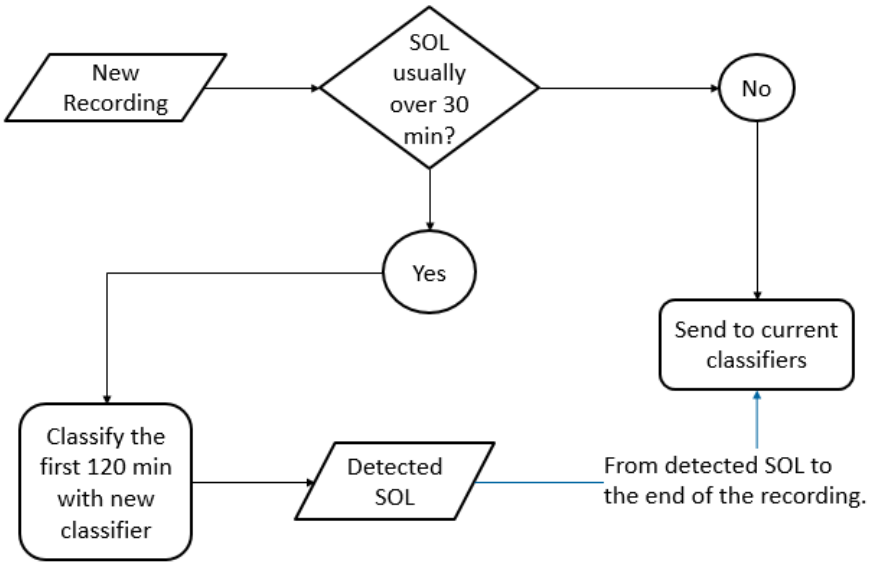


Figure 30.: Block diagram of logical steps and decisions regarding the implementation of the invented model for *sleep/wake* classification.

It is suggested that the subject completes a questionnaire on which the usual length of his/her SOL period is self-evaluated to fit one of the following categories: under or above

30 min. If it is the case when the subject usually does not reveal having delayed SOL (first category), then, most likely, current classification methods are adequate for the task of *sleep/wake* detection and SOL identification.
On the other hand, if the subject usually suffers from extended SOL, it is proposed that the classification regarding the beginning of the recordings (first 120 min) is performed separately from the classification over the remaining length of the recording.
In order to test the performance of the suggested model, the testing set presented on table 13 of chapter 5 was utilized. For subjects with $SOL_{GT} \geq 30$ min, the classification framework presented on figure 31 was adopted.

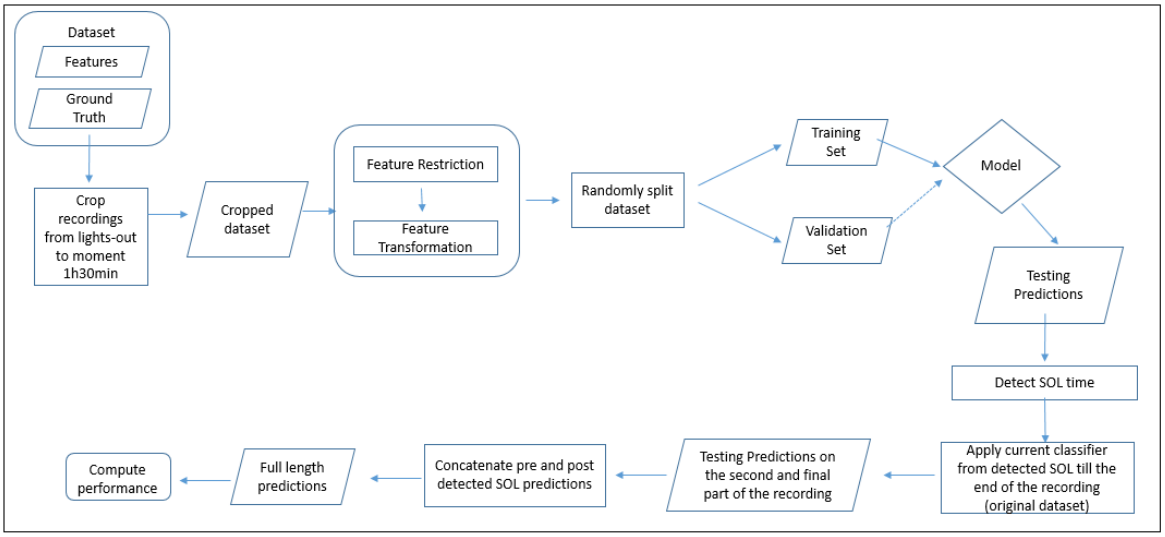


Figure 31.: High-level block diagram representing the reasoning process suggested for *sleep/wake* detection on subjects with delayed SOL_{GT}.

The reasoning includes:

1. Training a *sleep/wake* classifier [1] to recognize *wake* before SOL;
2. Applying that classifier to recordings clipped from the moment of lights off until the following 120 min;
3. Based on the *sleep/wake* epoch scoring for that period, compute SOL_d;
4. From the moment of SOL_d, until the end of each recording, apply a different classifier (based on a different set of features), trained to recognize *wake* on the overall length of the recordings.

It should be noted that each of the classifiers require specific sets of features. For the first one, the feature set will typically be more adequate to recognize periods of *wake* for which

1 The classifier must be trained with specific characteristics as described on chapter 5.

subjects might be lying still, while the second will typically be more adequate to recognize sudden changes which are typical of brief awakenings. Accordingly, FR was performed separately [2].

After obtaining the two feature sets, both classifiers were trained (each with example data of the correspondent period). Regarding the examples of data from subjects with delayed SOL_{GT} available (N=48), the features marked as included for model M_2 on table 14 were used for the first phase of the classification. The features utilized to train the second classifier are listed on table 28, in the appendix B.

Any classifier described in literature can be used. Notwithstanding, the results presented in the following section refer to the use of a Bayesian LD classifier [50] (for both periods of classification).

6.2 RESULTS

Table 16 and 17 discriminates the means and standard deviations values on the performance measures obtained from SOL detection and overall *sleep/wake* scoring statistics, for subjects with delayed SOL_{GT} and for the complete testing set, respectively.

Table 16.: Mean, standard deviation and range values regarding SOL-detection and overall classification performance obtained with the new and current *sleep/wake* detection techniques, for recording of subjects with delayed SOL_{GT} (N=48). For some metrics, the pooled values are also stated (before the mean and standard deviation values, which are between brackets). The significance of the difference between the results obtained with new approaches andcurrent classificationmethods, regarding L_1 error, κ coefficient, sensitivity, specificity, accuracy, precision and FPR, were calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Method	**New**		**Current**	
Parameter	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**
SOL_d **(min)**	34.02 ± 12.61	$17.00 - 88.00$	27.00 ± 18.07	$0.50 - 88.00$
SOL_{GT} **(min)**	52.58 ± 26.18	$30.00 - 126.00$	52.58 ± 26.18	$30.00 - 126.00$
L_1 **error (min)**	$\mathbf{20.04 \pm 23.42}^{**}$	$0.00 - 88.00$	$\mathbf{27.27 \pm 26.76}$	$0.50 - 104.00$
Bias (min)	$\mathbf{(-18.56) \pm 24.63}$	$\mathbf{(-88.00) - 8.50}$	$\mathbf{(-25.58) \pm 28.41}$	$\mathbf{(-104.00) - 35.00}$
κ	$\mathbf{0.64(0.65 \pm 0.16)}^{***}$	$0.12 - 0.87$	$\mathbf{0.60(0.61 \pm 0.16)}$	$0.06 - 0.87$
Sensitivity	$\mathbf{0.67(0.72 \pm 0.18)}^{***}$	$0.23 - 0.97$	$\mathbf{0.62(0.66 \pm 0.19)}$	$0.13 - 0.96$
Specificity	$0.94\ (0.93 \pm 0.05)^{**}$	$0.81 - 1.00$	$0.94\ (0.94 \pm 0.05)$	$0.82 - 1.00$
Accuracy	$0.87\ (0.87 \pm 0.10)^{***}$	$0.47 - 0.96$	$0.86\ (0.86 \pm 0.11)$	$0.32 - 0.97$
Precision	$0.80\ (0.81 \pm 0.14)^{*}$	$0.53 - 1.00$	$0.79\ (0.80 \pm 0.14)$	$0.52 - 1.00$
FPR	$0.06\ (0.06 \pm 0.05)^{**}$	$0.00 - 0.18$	$0.06\ (0.06 \pm 0.05)$	$0.00 - 0.18$

For ease of comparison, figures 32 and 33 illustrate the Bland-Altman plots on bias estimation resulting from SOL detection, using the methods described above, in section 6.1,

2 More information on the processes of FS for the first and second classifiers are available in chapter 5 and section 3.1.3, respectively.

Table 17.: Mean, standard deviation and range values regarding SOL-detection and overall classification performance obtained with the new and current *sleep/wake* detection techniques, for the complete heterogeneous testing set, of 170 recordings. For some metrics, the pooled values are also stated (before the mean and standard deviation values, which are between brackets). The significance of the difference between the results obtained with new approaches and current classification methods, regarding L_1 error, κ coefficient, sensitivity, specificity, accuracy, precision and FPR, were calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Method	**New**		**Current**	
Parameter	$\bar{x} \pm \sigma$	**Range**	$\bar{x} \pm \sigma$	**Range**
SOL_d **(min)**	20.82 ± 12.05	$0.50 - 88.00$	18.62 ± 12.79	$0.50 - 88.00$
SOL_{GT} **(min)**	24.47 ± 23.17	$1.50 - 126.00$	24.47 ± 23.17	$1.50 - 126.00$
L_1 **error (min)**	$\mathbf{9.26 \pm 14.56}^{***}$	$0.00 - 88.00$	$\mathbf{11.49 \pm 17.78}$	$0.00 - 104.00$
Bias (min)	$\mathbf{(-3.65) \pm 16.87}$	$\mathbf{(-88.00) - 18.00}$	$\mathbf{(-5.85) \pm 20.36}$	$\mathbf{(-104.00) - 35.00}$
κ	$\mathbf{0.61(0.59 \pm 0.16)}^{***}$	$0.05 - 0.87$	$\mathbf{0.59(0.58 \pm 0.16)}$	$0.06 - 0.87$
Sensitivity	$\mathbf{0.68(0.73 \pm 0.17)}^{***}$	$0.14 - 0.97$	$\mathbf{0.66(0.70 \pm 0.18)}$	$0.13 - 0.97$
Specificity	$0.93\ (0.93 \pm 0.06)^{**}$	$0.73 - 1.00$	$0.93\ (0.93 \pm 0.06)$	$0.73 - 1.00$
Accuracy	$0.88\ (0.88 \pm 0.07)^{***}$	$0.47 - 0.97$	$0.87\ (0.87 \pm 0.08)$	$0.32 - 0.97$
Precision	$0.69\ (0.68 \pm 0.20)^{*}$	$0.10 - 0.99$	$0.69\ (0.68 \pm 0.20)$	$0.10 - 1.00$
FPR	$0.07\ (0.07 \pm 0.06)^{**}$	$0.00 - 0.27$	$0.07\ (0.07 \pm 0.06)$	$0.00 - 0.27$

and with current methodologies [3], for subjects with delayed SOL_{GT} (over 30 min). On figure 32 the values presented correspond to the results on SOL detection obtained from the first classifier.

Figures 34 and 35 refer to the same comparison (new method vs. current method), but regarding all subjects on the testing set, with the goal of addressing the influence of the changes obtained for the delayed SOL_{GT} group, over the complete validation set, of 170 subjects. Logically, between figure 34 and 35, only the SOL_d results of recordings with delayed SOL_{GT} differ.

6.3 DISCUSSION

Table 16 shows that, for any of the performance metrics presented, there are significant improvements when using the new *sleep/wake* classification method instead of the current one. Specifically, there was an increase of overall sensitivity from 62% to 67%, which indicates the meliorated capacity of the new classification method to discriminate the positive class, *wake*. Also, the κ coefficient increased significantly, from pooled 0.60 to 0.64, which suggests better agreement.

Table 17 shows that, not only were the improvements significant for the delayed SOL_{GT} group, but also the significance still applies, when increasing the sample number for 48 to 170 recordings, even if the added 122 recordings were not subjected to any alterations

3 Current classification methodologies are described in section 3.2.

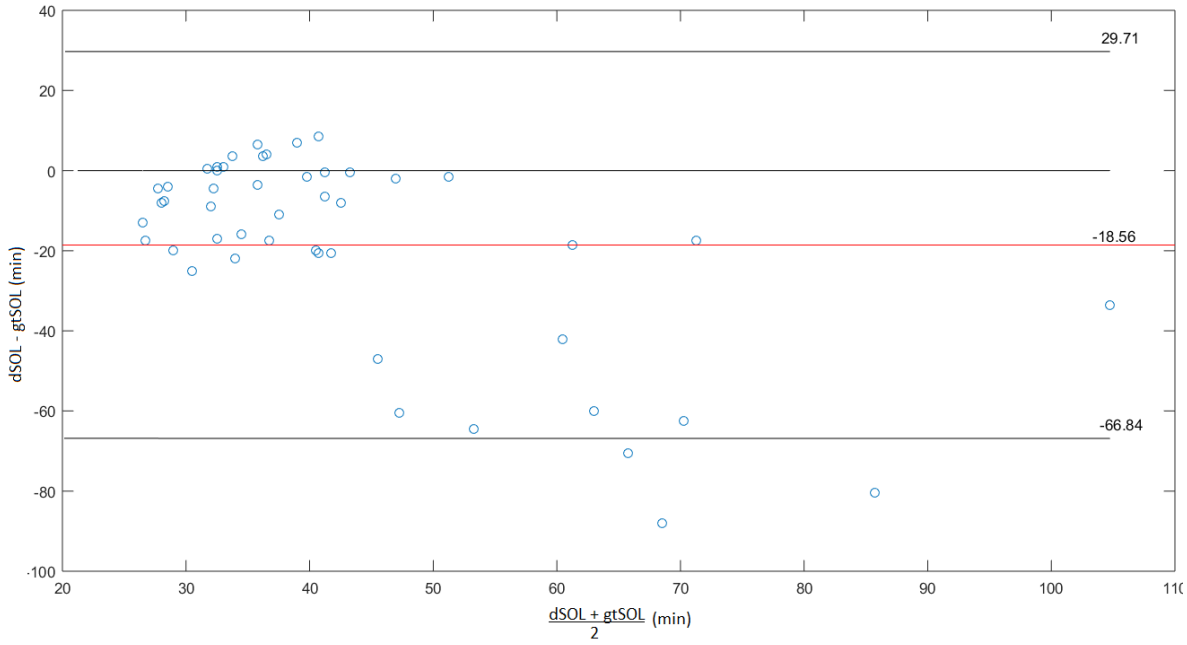


Figure 32.: Bland-Altman plot representation of bias estimation on SOL detection obtained with the new classification technique for subjects with $SOL_{GT} \geq 30$ min.

regarding classification, thus, presenting the same values on performance parameters as in previous current classification results, described in section 3.2.
Regarding specifically the results obtained on SOL-detection for subjects with delayed SOL_{GT}, from the comparison of the two Bland-Altman plots on figures 32 and 33, it can be seen that, even though underestimation of SOL is still happening, when applying the new technique, the mistakes being made are of a much smaller magnitude than what is observed with current classification methods. One obvious indicator of that is the average bias that was reduced from 25.58 min of underestimation to 18.58 min of underestimation of SOL. There is also a large reduction on the standard deviation of the bias. The interval ranging between the upper and lower limits of confidence was reduced by approximately 31 min.
Figure 34 makes it even more evident that, even though the SOL detection improvement occurred only for 48 subjects, it has had a major impact on the general SOL-detection grouped performance. The average bias is now closer to zero minutes and the standard deviation value associated to it was reduced, revealing the elimination of the most critical outliers.
Regarding results on overall statistical performance, contained on tables 16 and 17, the improvement is not so obvious.

This is due to the fact the the majority part of the epoch *sleep/wake* scoring is identical to the one obtained from current classification methods. Specifically, for the delayed SOL_{GT} group, the scoring is provided by current classifiers from the moment of SOL_d (which is

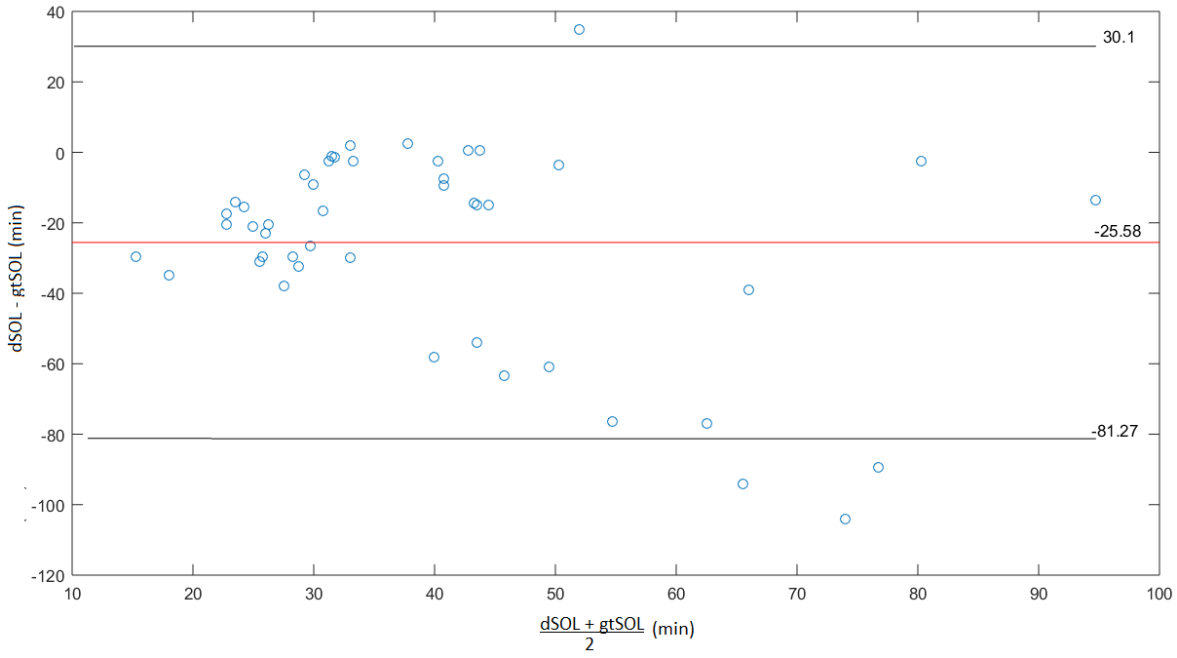


Figure 33.: Bland-Altman plot representation of bias estimation on SOL-detection obtained with current classification methods, for subjects with SOL_{GT} over 30 min (N=48).

somewhere within the first 120 min) until the end of the recordings; and, for the remaining 122 recordings, the scoring is exactly the same since the moment of lights off until the end of the recordings. Nevertheless, after performing a two-tailed Wilcoxon-signed rank test, the differences on all the statistics metrics of performance proved to be significant, not only for the group that was subjected to alterations, but also its influence on the complete group of subjects positively influenced the results.

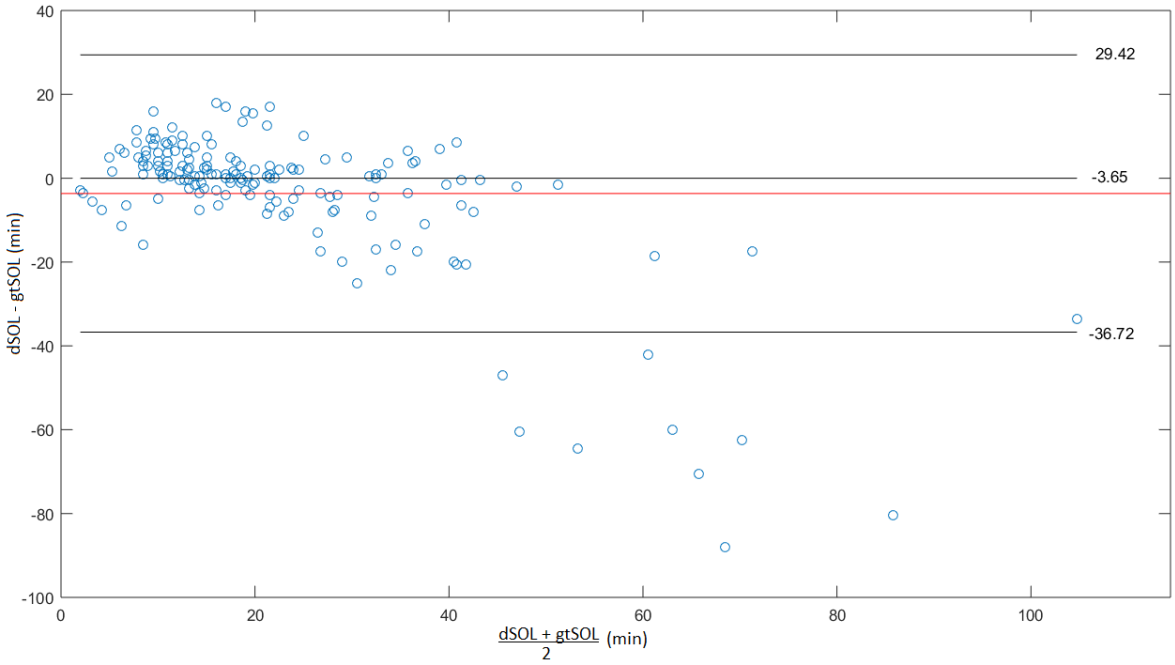


Figure 34.: Bland-Altman plot representation of bias estimation on SOL detection obtained with the new classification technique for the complete testing set.

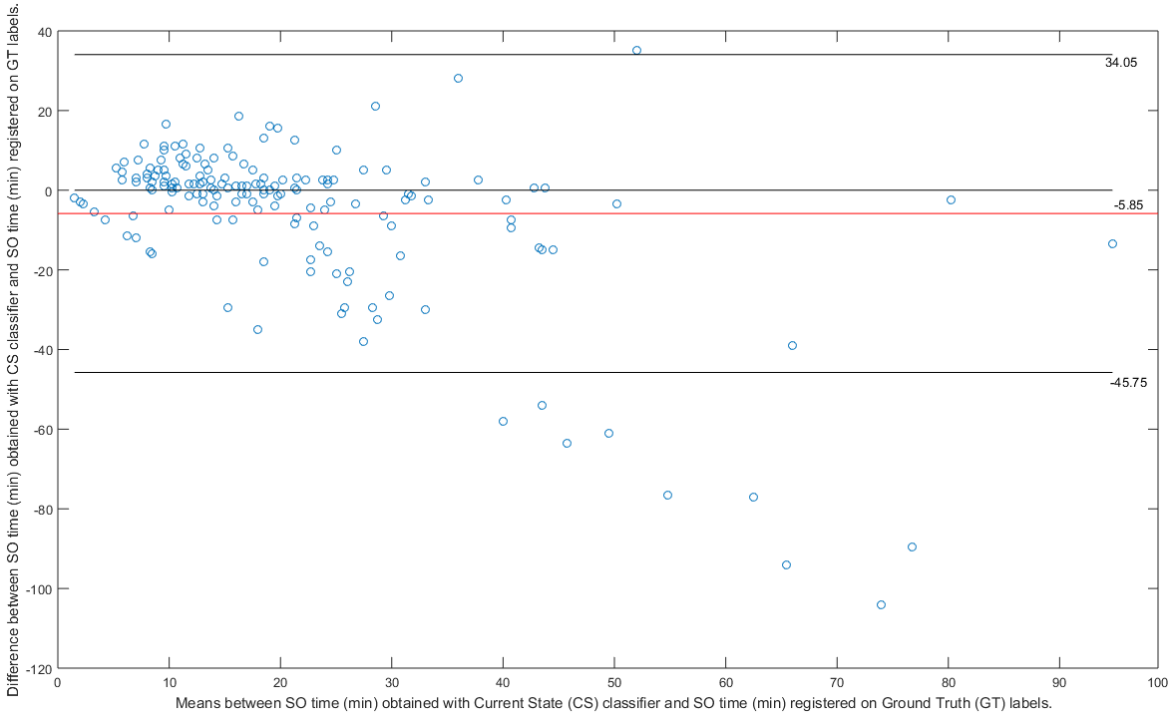


Figure 35.: Bland-Altman plot representation of bias estimation on SOL-detection obtained with current classification methods, for the complete testing set (N=170).

7

CONCLUSIONS

7.1 CONCLUSION

This work addressed the performance of unobtrusive modalities for *sleep/wake* classification and SOL detection, based on cardiorespiratory features.
It was shown that the characteristics that distinguish *wake* from *sleep* are different before and after initiating sleep, specially for subjects that take longer than 30 minutes to fall asleep, probably due to their usual behavior during the period of SOL. These are subjects who express typical characteristics of sleep (forcing their bodies to lie still and calming respiration) for long periods of time, while still awake. In these situations, even though actigraphy has been indicated by the AASM [64] as a suitable method to assist in the evaluation of sleep, actigraphy-based classification is not adequate, since it relies to a great extent on the activity counts in a given epoch.
Hence, in this work we propose a new *sleep/wake* detection reasoning, for subjects with extended SOL periods, which comprises two steps: firstly, it recognizes *wake* during SOL and, secondly, it recognizes *wake* after SOL. While the second step of the classification was performed as in current methods, the first one makes use of a specific set of features, which depend on the usual estimated SOL time of each subject. For subjects who do not usually experience delayed SOL, *sleep/wake* detection was performed by current classifiers, over the complete length of the recordings.
The new classification approach was evaluated in an hold-out scheme, on a data set constituted by 339 PSG recordings of healthy subjects (having used 169 recordings for training and 170 recordings for testing). For subjects with SOL periods over 30 minutes, a significant improvement in all performance metrics was obtained. Specifically, it achieved an increased Cohen's Kappa coefficient ($\kappa = 0.64$) compared with current classification methods ($\kappa = 0.60$), and an increased sensitivity from overall 62% to 67%, confirming the meliorated capacity of new methods to detect the positive class (*wake*). Moreover, regarding L_1 error on SOL detection, when performing the classification as suggested, there was a significant decrease from 27.27 ± 26.76 min to 20.04 ± 23.42 min. The achieved results also demonstrated to positively influence overall performance of sleep detection algorithms, when in the case

of an heterogeneous population, regarding SOL time.
In summary, the proposed method significantly outperforms current classifiers, without introducing much complexity in the process. The progresses made regarding SOL detection can be important in avoiding underestimating symptoms of sleep disorders, such as insomnia, during medical diagnosis.

7.2 PROSPECT FOR FUTURE WORK

As future work, it would be advantageous to implement a method that would pre-select the subjects for the correct algorithm, according to estimated SOL, given the subjectivity of the proposed self-evaluation of this metric. We have tried to implement such a method, firstly using machine learning techniques and, secondly, by addressing the correlation between SOL_{GT} or SOL_d and the quantity of some features, such as the actigraphy feature. However, the former led to poor validation results (probably because the feature space available is not adequate for that task, but for the task of discriminating sleep stages), besides being complex and time-consuming. The second approach was also inconclusive.
It would be interesting to evaluate if the proposed methodology allows for comparable performance on subjects with sleep disorders, such as sleep apnea and/or insomnia. If successful, this method would be a strong candidate to implement on sleep home monitoring and medical diagnosis practices, considering performance as well as its unobtrusive potential, associated to low costs and complexity.
Also, the investigation of new features to better express *wake* before SOL, in replacement of actigraphy-based features, could possibly lead to enhancements in the classification performance. Likewise, research could be done in order to probe whether the classification method for the second part of the recordings could be further improved. Due to lack of time, this possibility was not investigated in this work. Hence, current classification methods were used for this task.
Another possible direction of future work is to address the influence of the first night effect on the results presented in this report regarding SOL detection and *sleep/wake* detection during this period.

BIBLIOGRAPHY

[1] American Academy of Sleep Medicine. Quality measures. http://www.aasmnet.org/, 2016.

[2] L. Mastin. How sleep works. http://www.howsleepworks.com, 2013.

[3] K. Morton. Why learn about sleep? http://www.end-your-sleep-deprivation.com/, 2015.

[4] World Association of Sleep Medicine. Advancing sleep health worldwide. http://www.wasmonline.org/, 2016.

[5] S.J. Redmond and C. Heneghan. Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea. *Biomedical Engineering, IEEE Transactions on*, 53(3):485–496, 2006.

[6] P.A. Estévez, C.M. Held, C.A. Holzmann, C.A. Perez, J.P. Pérez, J. Heiss, M. Garrido, and P. Peirano. Polysomnographic pattern recognition for automated classification of sleep-waking states in infants. *Medical and Biological Engineering and Computing*, 40(1):105–113, 2002.

[7] S. Devot, R. Dratwa, and E. Naujokat. Sleep/wake detection based on cardiorespiratory signals and actigraphy. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 5089–5092. IEEE, 2010.

[8] National Sleep Foundation. Sleeptionary - definitions of common sleep terms. https://sleepfoundation.org/sleeptionary, 2016.

[9] Philips Healthcare. Philips respironics actiwatch. http://www.actiwatch.respironics.com, 1999.

[10] W. Karlen, C. Mattiussi, and D. Floreano. Improving actigraph sleep/wake classification with cardio-respiratory signals. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 5262–5265. IEEE, 2008.

[11] S.J. Redmond, P. de Chazal, C. O'Brien, S. Ryan, W.T. McNicholas, and C. Heneghan. Sleep staging using cardiorespiratory signals. *Somnologie-Schlafforschung und Schlafmedizin*, 11(4):245–256, 2007.

[12] T. Willemen, D. Van Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S. Van Huffel, B. Haex, and J. Vander Sloten. An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification. *Biomedical and Health Informatics, IEEE Journal of*, 18(2):661–669, 2014.

[13] A.A. Borbély. A two process model of sleep regulation. *Human neurobiology*, 1982.

[14] A.A. Borbély and P. Achermann. Sleep homeostasis and models of sleep regulation. *Journal of biological rhythms*, 14(6):559–570, 1999.

[15] J.R.L. Schwartz and T. Roth. Neurophysiology of sleep and wakefulness: basic science and clinical implications. *Current neuropharmacology*, 6(4):367–378, 2008.

[16] Harvard Medical School. Healthy sleep. `http://healthysleep.med.harvard.edu`, 2008.

[17] T. Åkerstedt. Shift work and disturbed sleep/wakefulness. *Occupational Medicine*, 53(2):89–94, 2003.

[18] T.L. Lee-Chiong. *Sleep: a comprehensive handbook*. Wiley Online Library, 2006.

[19] M.H. Kryger, T. Roth, and W.C. Dement. *Principles and Practice of Sleep Medicine*. Elsevier, 5 edition, 2011.

[20] American Academy of Sleep Medicine, Conrad Iber, et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, 2007.

[21] A. Rechtschaffen and A. Kales. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. 1968.

[22] S. Laureys, O. Gosseries, and G. Tononi. *The Neurology of Consciousness: Cognitive Neuroscience and Neuropathology*. Elsevier Science, 2015.

[23] Sleepdex. Stages of sleep. `http://www.sleepdex.org/stages.htm`, 2016.

[24] Z. Shinar, S. Akselrod, Y. Dagan, and A. Baharav. Autonomic changes during wake–sleep transition: A heart rate variability based approach. *Autonomic Neuroscience*, 130(1):17–27, 2006.

[25] C. Blume, R. del Giudice, M. Wislowska, J. Lechinger, and M. Schabus. Across the consciousness continuum—from unresponsive wakefulness to sleep. *Frontiers in human neuroscience*, 9, 2015.

[26] C.A. Kushida, M.R. Littner, T. Morgenthaler, C.A. Alessi, D. Bailey, J. Coleman Jr, L. Friedman, M. Hirshkowitz, S. Kapen, M. Kramer, et al. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. *Sleep*, 28(4):499–521, 2005.

[27] S. Koos. Polysomnography. `http://www.chicagosleepapneasnoring.com`, 2011.

[28] Nicola Cellini. The effects of sleep on autonomic regulation, cardiovascular activity, and cognitive processing.

[29] X. Long. *On the Analysis and Classification of Sleep Stages from Cardiorespiratory Activity*. PhD thesis, Technische Universiteit Eindhoven, 2015.

[30] N.J. Douglas, D.P. White, C.K Pickett, J.V. Weil, and C.W. Zwillich. Respiration during sleep in normal man. *Thorax*, 37(11):840–844, 1982.

[31] D.J. Buysse, C.F. Reynolds, T.H. Monk, S.R. Berman, and D.J. Kupfer. The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry research*, 28(2):193–213, 1989.

[32] D. Crean. *Sleep: Rock Thy Brain*. CreateSpace Independent Publishing Platform, 2015.

[33] M.A. Carskadon, W.C. Dement, M.M. Mitler, T. Roth, P.R. Westbrook, and S. Keenan. Guidelines for the multiple sleep latency test (mslt): a standard measure of sleepiness. *Sleep*, 9(4):519–524, 1986.

[34] H.W. Agnew, W.B. Webb, and R.L. Williams. The first night effect: An eeg study of sleep. *Psychophysiology*, 2(3):263–266, 1966.

[35] G. Klosh, B. Kemp, T.H. Penzel, A. Schlögl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeithofer, B. Saletu, W.M. Herrmann, et al. The siesta project polygraphic and clinical database. *Engineering in Medicine and Biology Magazine, IEEE*, 20(3):51–57, 2001.

[36] P. Fonseca, X. Long, M. Radha, R. Haakma, R.M. Aarts, and J. Rolink. Sleep stage classification with ecg and respiratory effort. *Physiological measurement*, 36(10):2027, 2015.

[37] P. Fonseca, R.M. Aarts, X. Long, J. Rolink, and S. Leonhardt. Estimating actigraphy from motion artifacts in ecg and respiratory effort signals. *Physiological Measurement*, 37(1):67, 2016.

[38] R.A. Rhoades and D.R. Bell. *Medical phisiology: Principles for clinical medicine*. Lippincott Williams & Wilkins, 2012.

[39] Task Force of the European Society of Cardiology et al. Heart rate variability standards of measurement, physiological interpretation, and clinical use. *Eur Heart J*, 17:354–381, 1996.

[40] M. Basner, B. Griefahn, U. Muller, G. Plath, and A. Samel. An ecg-based algorithm for the automatic identification of autonomic activations associated with cortical arousal. *SLEEP-NEW YORK THEN WESTCHESTER-*, 30(10):1349, 2007.

[41] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R.M. Aarts. Using dynamic time warping for sleep and wake discrimination. In *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, pages 886–889. IEEE, 2012.

[42] X. Long, J. Yang, T. Weysen, R. Haakma, J. Foussier, P. Fonseca, and R.M. Aarts. Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. *Physiological measurement*, 35(12):2529, 2014.

[43] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R.M. Aarts. Sleep and wake classification with actigraphy and respiratory effort using dynamic warping. *Biomedical and Health Informatics, IEEE Journal of*, 18(4):1272–1284, 2014.

[44] X. Long, P. Fonseca, R.M. Aarts, R. Haakma, and J. Foussier. Modeling cardiorespiratory interaction during human sleep with complex networks. *Applied Physics Letters*, 105(20):203701, 2014.

[45] Y. Jie. Feature extraction for deep sleep detection.

[46] D. Cysarz, H. Bettermann, S. Lange, D. Geue, and P. Van Leeuwen. A quantitative comparison of different methods to detect cardiorespiratory coordination during nighttime sleep. *Biomed Eng Online*, 3(1):44, 2004.

[47] H. Bettermann, D. Cysarz, and P. Van Leeuwen. Detecting cardiorespiratory coordination by respiratory pattern analysis of heart period dynamics—the musical rhythm approach. *International Journal of Bifurcation and Chaos*, 10(10):2349–2360, 2000.

[48] J. Foussier, P. Fonseca, X. Long, and S. Leonhardt. Automatic feature selection for sleep/wake classification with small data sets. In *BIOINFORMATICS*, pages 178–184, 2013.

[49] M.A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[50] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[51] A.K. Jain. Advances in statistical pattern recognition. In *Pattern recognition theory and applications*, pages 1–19. Springer, 1987.

[52] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

[53] H. Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361, 1920.

[54] D.G. Altman and J.M. Bland. Measurement in medicine: the analysis of method comparison studies. *The statistician*, pages 307–317, 1983.

[55] John W Tukey. Exploratory data analysis. 1977.

[56] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[57] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R.M. Aarts. Analyzing respiratory effort amplitude for automated sleep stage classification. *Biomedical Signal Processing and Control*, 14:197–205, 2014.

[58] A.J. Viera, J.M. Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[59] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[60] J. Paquet, A. Kawinska, and J. Carrier. Wake detection capacity of actigraphy during sleep. *SLEEP-NEW YORK THEN WESTCHESTER-*, 30(10):1362, 2007.

[61] K. Pearson. Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367):489–498, 1896.

[62] T. Penzel, J.W. Kantelhardt, L. Grote, J. Peter, and A. Bunde. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *Biomedical Engineering, IEEE Transactions on*, 50(10):1143–1151, 2003.

[63] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R.M. Aarts. Sleep and wake classification with actigraphy and respiratory effort using dynamic warping. *Biomedical and Health Informatics, IEEE Journal of*, 18(4):1272–1284, 2014.

[64] American Academy of Sleep Medicine. Sleep deprivation. `www.aasmnet.org/`, 2016.

[65] A. Noviyanto, S.M. Isa, I. Wasito, A.M. Arymurthy, et al. Selecting features of single lead ecg signal for automatic sleep stages classification using correlation-based feature subset selection. *IJCSI International Journal of Computer Science Issues*, 8(5), 2011.

[66] X. Long, R. Haakma, J. Rolink, P. Fonseca, and R.M. Aarts. Improving sleep/wake detection via boundary adaptation for respiratory spectral features. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 374–377. IEEE, 2015.

A

TIME WINDOW ANALYSIS FOR SOL DETECTION

With the goal of selecting the most appropriate time window, for increased performance on SOL-detection, the night recordings were clipped in different time ranges.
The data set utilized for this analysis is described on table 13 and the method of classification applied is described on figure 25 (both included in section 5.1, chapter 5).

A.1 RESULTS

Here are included the performance results on the testing set regarding *sleep/wake* detection on the beginning of the recordings:

- Table 18 contains the results regarding kappa and SOL detection performance metrics for a time window of 31 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 36;
- Table 19 contains the results regarding kappa and SOL detection performance metrics for a time window of 40 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 37;
- Table 20 contains the results regarding kappa and SOL detection performance metrics for a time window of 50 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 38;
- Table 21 contains the results regarding kappa and SOL detection performance metrics for a time window of 60 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 39;
- Table 22 contains the results regarding kappa and SOL detection performance metrics for a time window of 70 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 40;

- Table 23 contains the results regarding kappa and SOL detection performance metrics for a time window of 80 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 41;
- Table 24 contains the results regarding kappa and SOL detection performance metrics for a time window of 90 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 42;
- Table 25 contains the results regarding kappa and SOL detection performance metrics for a time window of 100 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 43;
- Table 26 contains the results regarding kappa and SOL detection performance metrics for a time window of 110 min.A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 44;
- Table 27 contains the results regarding kappa and SOL detection performance metrics for a time window of 120 min. A comparison regarding bias estimation of the different classification approaches, in this window is presented on figure 45.

Table 18.: Performance measures obtained for the first 31 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects, regarding the training set. The feature set for FR was altered between: CFS FS (std); union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**31 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.60	0.04	0.38	0.44	0.00	0.00	0.43	0.05	0.03	0.58
d L_1 **Error (min)**	12.29±18.97	23.57±23.13***	14.75±19.89***	13.90±19.69**	24.09±23.18***	24.09±23.18***	13.73±19.18**	22.76±22.86***	23.49±23.27***	11.40±17.92
Bias (min)	(-5.77)±21.87	(-23.47)±23.23	(-8.91)±23.12	(-7.92)±22.78	(-24.09)±23.18	(-24.09)±23.18	(-8.25)±22.11	(-22.67)±22.95	(-23.43)±23.32	(-5.42)±20.55
SOL_d **(min)**	18.82±7.85	1.12±1.52	15.68±0.25	16.68±7.51	0.50	0.50	16.35±8.73	1.92±2.75	1.16±1.43	19.17±12.45
SOL_{GT} **(min)**	24.59±23.18									
N	169 recordings									

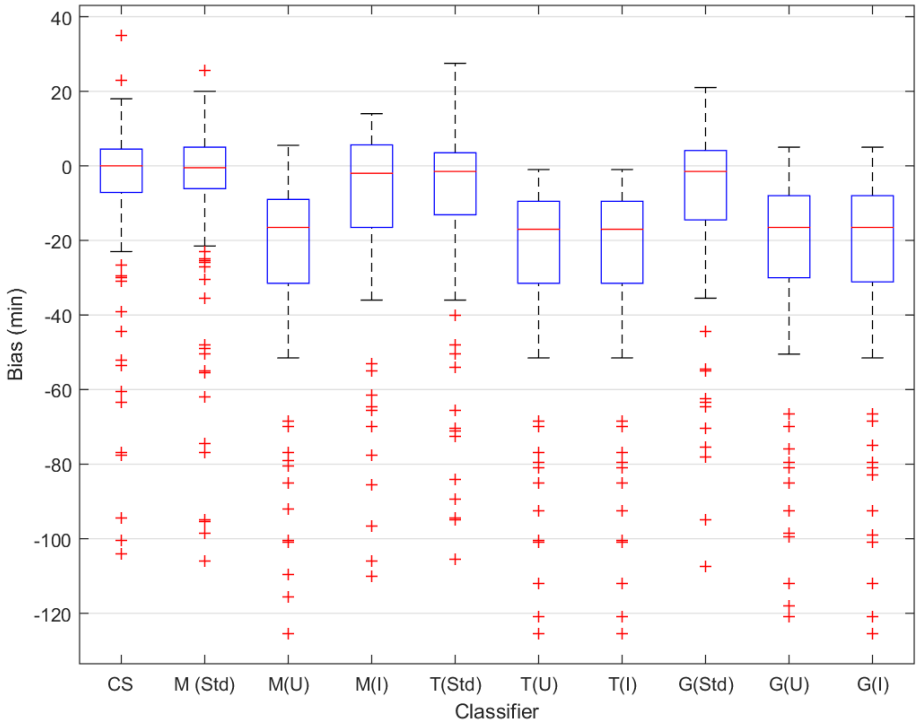


Figure 36.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 31 min.

Table 19.: Performance measures obtained for the first 40 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects, regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**40 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.35	0.36	0.00	0.36	0.38	0.48	0.36	0.49	0.48	0.58
L_1 **Error (min)**	17.14±20.52***	17.75±21.41***	23.97±23.17***	17.42±21.19***	19.40±21.72***	14.61±18.28**	17.03±20.86***	13.46±19.18*	13.54±17.98***	11.29±17.55
Bias (min)	(-11.18)±24.31	(-13.87)±24.12	(-23.97)±23.17	(-13.08)±24.12	(-17.49)±23.29	(-8.03)±22.00	(-10.66)±4.75	(-6.74)±22.46	(-4.22)±22.14	(-5.15)±20.24
SOL_d **(min)**	13.29±13.97	10.60±12.29	0.50±0.00	11.39±12.34	6.98±10.40	16.44±13.11	13.81±13.83	17.73±7.33	20.25±5.65	19.32±16.65
SOL_{GT} **(min)**	24.47±23.17									
N	170 recordings									

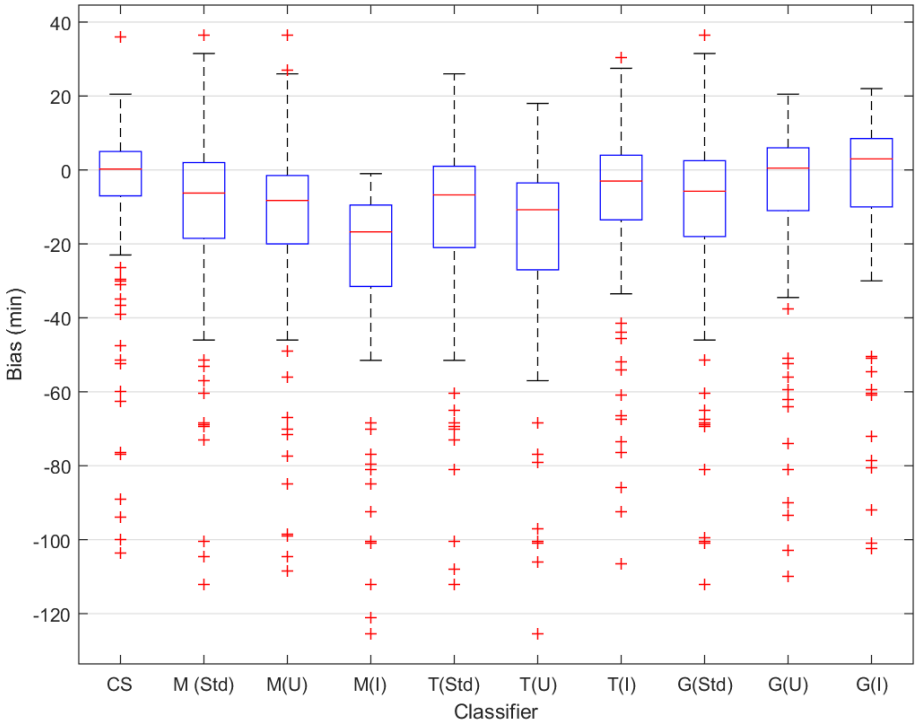


Figure 37.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 40 min.

Table 20.: Performance measures obtained for the first 50 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects, regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**50 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.40	0.43	0.43	0.34	0.42	0.50	0.49	0.43	0.43	0.58
L_1 Error (min)	16.59±20.87***	16.00±20.25***	16.29±19.86***	18.07±21.19***	18.09±21.68***	14.58±18.73**	15.21±18.98***	15.93±20.22***	16.46±19.77***	11.49±17.78
Bias (min)	(-11.42)±24.10	(-9.20)±24.13	(-10.13)±23.62	(-14.23)±23.95	(-13.04)±25.07	(-6.86)±22.74	(-4.63)±23.90	(-9.26)±24.04	(-9.28)±24.02	(-5.85)±20.36
SOL_d (min)	13.05±13.18	15.27±13.56	14.34±13.77	10.24±11.90	11.44±14.17	17.61±15.22	19.84±13.14	15.21±13.15	15.19±14.41	18.59±12.76
SOL_{GT} (min)	24.47±23.17									
N	170 recordings									

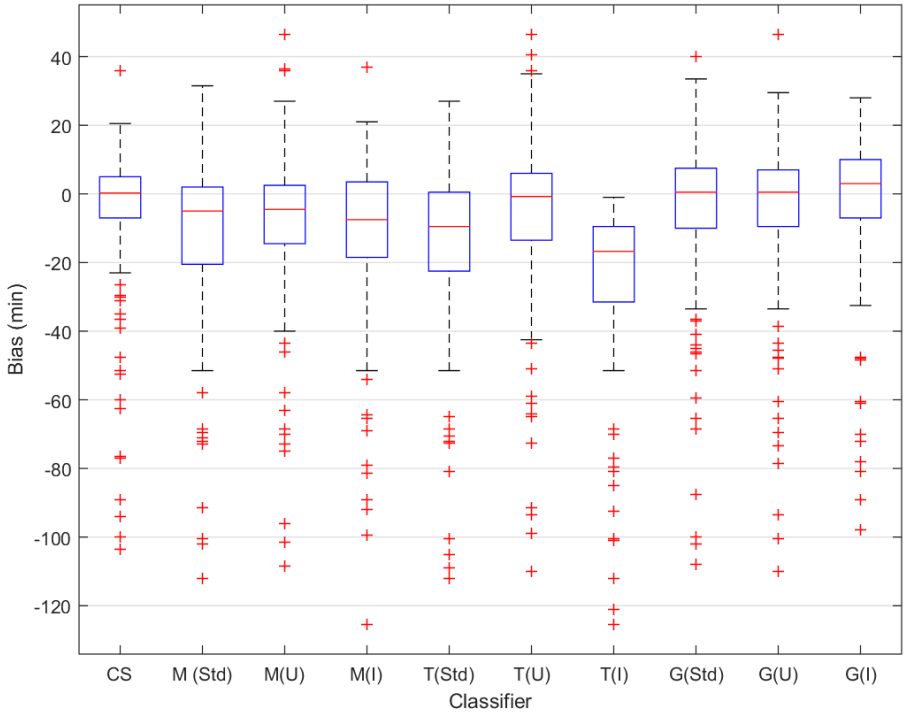


Figure 38.: Box-and-whisker plot, comparing the results on bias estimation of SO detection, with different classifiers, over a window of 50 min.

Table 21.: Performance measures obtained for the first 60 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**60 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.34	0.42	0.42	0.36	0.41	0.48	0.43	0.52	0.54	0.58
L_1 Error (min)	17.99±21.68***	15.95±20.63***	16.12±19.96***	17.54±21.15***	18.59±21.75***	15.56±19.75***	15.94±20.03***	13.34±19.09*	14.04±17.03**	11.49±17.78
Bias (min)	(-14.35)±24.25	(-10.72)±23.79	(-11.34)±23.03	(-13.79)±23.77	(-15.16)±24.28	(-10.77)±22.74	(-8.03)±24.33	(-5.96)±22.54	(-2.39)±21.96	(-5.85)±20.36
SOL_d (min)	10.12±12.08	13.75±12.40	13.13±13.04	10.68±12.07	9.31±12.63	13.70±12.77	16.44±13.33	18.51±8.54	22.08±10.32	18.59±12.76
SOL_{GT} (min)	24.47±23.17									
N	170 recordings									

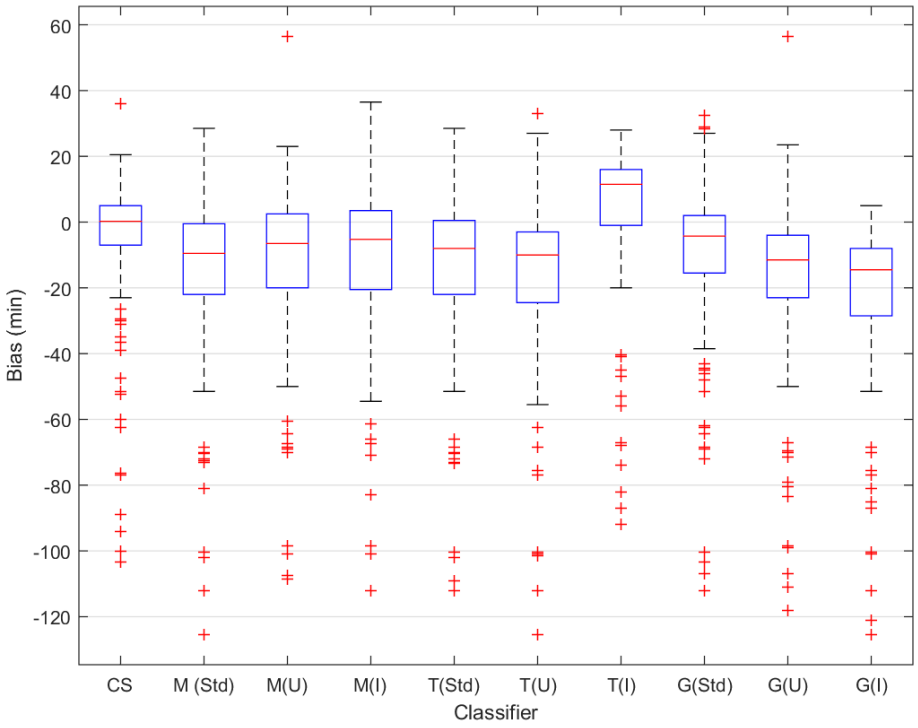


Figure 39.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 60 min.

Table 22.: Performance measures obtained for the first 70 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects regarding the training set. The feature set for FR was altered between: standard (std) feature selection; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**70 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.32	0.38	0.40	0.29	0.39	0.45	0.36	0.38	0.40	0.58
L_1 Error (min)	17.78±21.39***	16.10±20.42***	16.13±20.13***	19.01±22.60***	15.92±20.17***	15.72±19.77***	17.34±21.86***	16.15±21.05***	15.81±20.14***	11.49±17.78
Bias (min)	(-14.22)±23.92	(-11.21)±23.48	(-11.74)±22.99	(-17.04)±24.13	(-11.90)±22.79	(-10.76)±22.87	(-13.04)±24.68	(-11.26)±24.04	(-11.42)±22.93	(-5.85)±20.36
SOL_d (min)	10.25±11.77	13.26±13.28	12.74±12.44	7.43±10.09	12.57±12.73	13.71±13.06	11.43±12.29	13.21±12.36	13.05±12.65	18.59±12.76
SOL_{GT} (min)	24.47±23.17									
N	170 recordings									

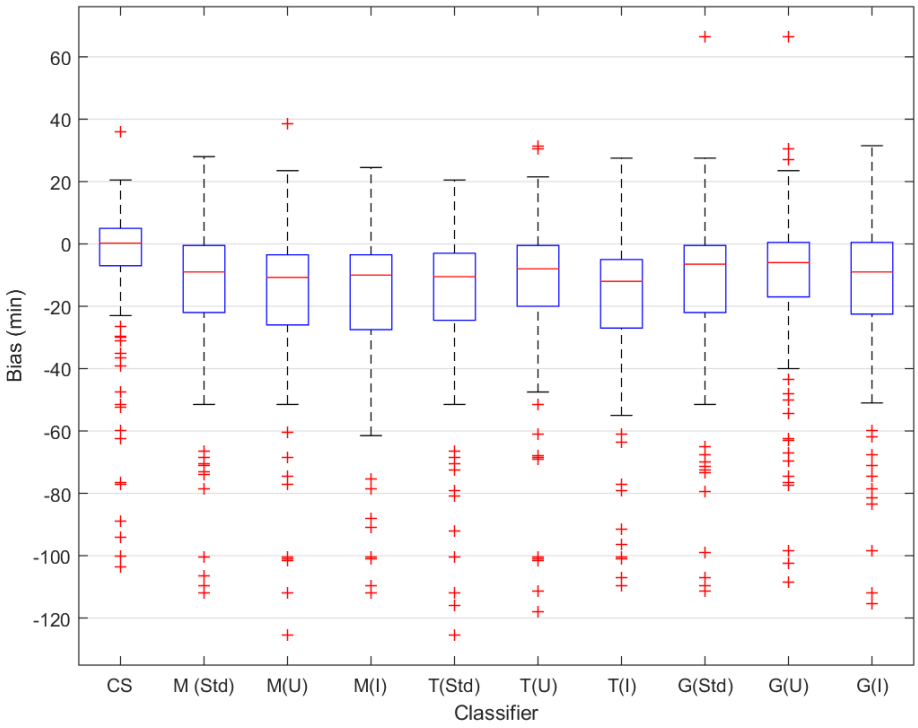


Figure 40.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 70 min.

Table 23.: Performance measures obtained for the first 80 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**80 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.30	0.34	0.35	0.29	0.36	0.42	0.61	0.51	0.53	0.58
L_1 Error (min)	18.08±21.25***	16.13±20.57***	16.13±20.20***	18.19±21.44***	16.06±20.37***	15.46±19.85***	12.10±18.46	17.96±21.09***	13.18±17.13**	11.49±17.78
Bias (min)	(-14.61)±23.78	(-11.01)±23.73	(-11.95)±22.94	(-14.67)±24.00	(-12.21)±22.91	(-10.92)±22.69	(-5.07)±21.49	(-14.14)±23.83	(-3.34)±21.38	(-5.85)±20.36
SOL_d (min)	9.86±11.50	13.46±13.43	12.52±12.41	9.80±11.52	12.26±12.44	13.55±12.76	19.40±10.85	10.33±12.66	21.13±13.41	18.59±12.76
SOL_{GT} (min)	24.47±23.17									
N	170 recordings									

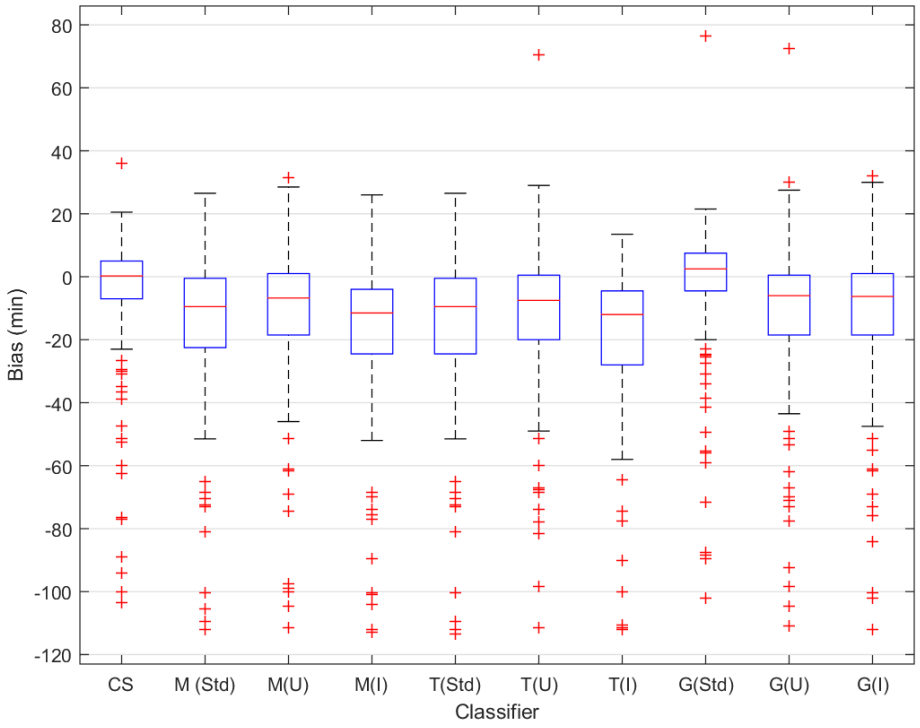


Figure 41.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 80 min.

Table 24.: Performance measures obtained for the first 90 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**90 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.25	0.32	0.32	0.25	0.30	0.38	0.29	0.44	0.51	0.58
L_1 **Error (min)**	18.05±21.20***	16.72±20.70***	16.46±20.73***	18.13±21.23***	16.79±21.02***	15.73±20.13***	18.24±21.58***	15.61±21.18***	13.62±19.05*	11.49±17.78
Bias (min)	(-14.46)±23.81	(-13.61)±22.89	(-13.40)±22.84	(-14.57)±23.83	(-12.66)±23.75	(-11.81)±22.67	(-14.58)±24.22	(-10.40)±24.18	(-8.54)±21.82	(-5.85)±20.36
SOL_d **(min)**	10.01±11.57	10.86±11.88	11.07±11.64	9.90±11.59	11.81±12.49	12.66±12.35	9.89±10.21	14.07±12.13	15.94±12.53	18.59±12.76
SOL_{GT} **(min)**	24.47±23.17									
N	170 recordings									

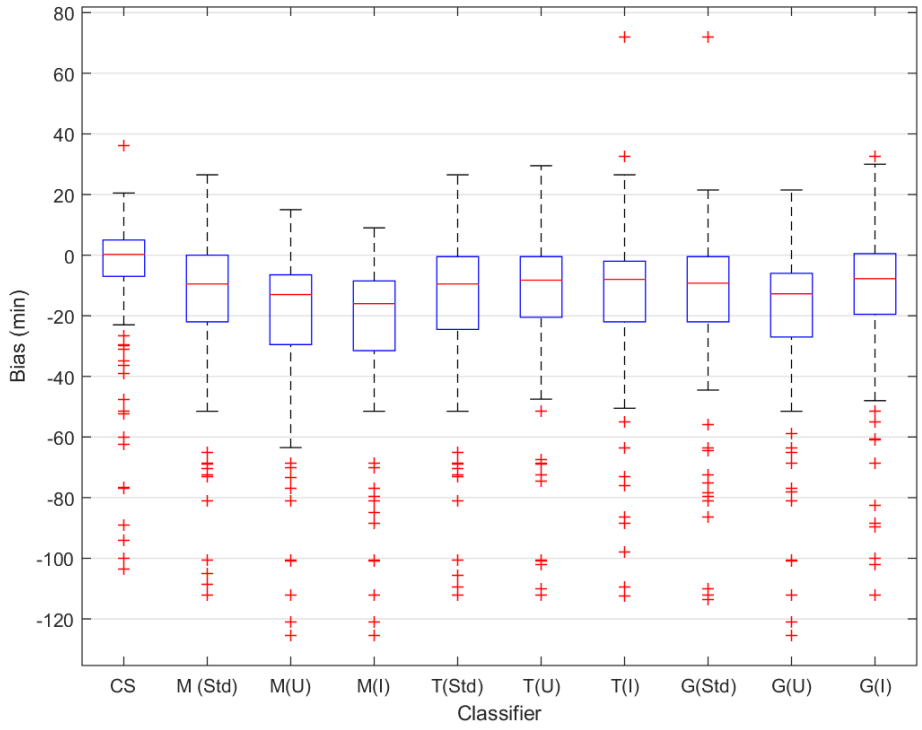


Figure 42.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 90 min.

Table 25.: Performance measures obtained for the first 100 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) was calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**100 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.63	0.33	0.35	0.30	0.35	0.42	0.31	0.33	0.50	0.58
L_1 **Error (min)**	11.83±15.11*	16.75±20.11***	16.52±19.96***	18.22±21.13***	16.29±20.04***	15.14±19.46***	17.28±20.98***	17.13±21.15***	13.32±18.00*	11.49±17.78
Bias (min)	0.47±19.21	(-13.43)±22.48	(-13.78)±21.95	(-13.52)±24.43	(-11.89)±22.94	(-10.75)±22.20	(-12.99)±23.89	(-14.14)±23.27	(-6.86)±21.34	(-5.85)±20.36
SOL_d **(min)**	24.94±13.78	11.04±12.47	10.69±12.85	10.95±14.13	12.58±15.20	13.72±14.74	11.48±13.26	10.33±12.45	17.61±14.34	18.59±12.76
SOL_{GT} (min)	24.47±23.17									
N	170 recordings									

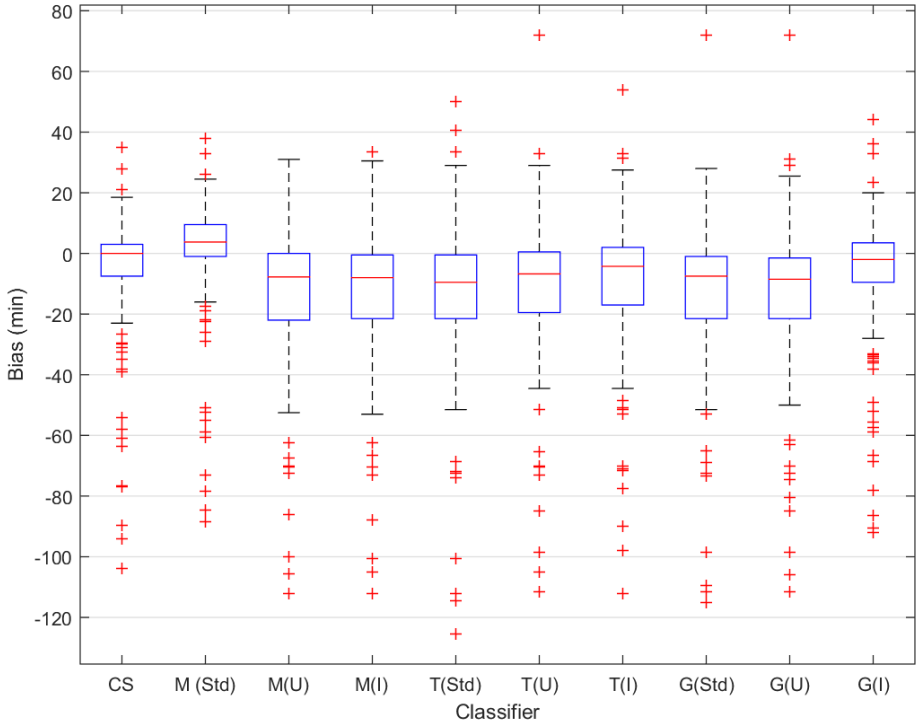


Figure 43.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 100 min.

Table 26.: Performance measures obtained for the first 110 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification results (CS) calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**110 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.63	0.38	0.34	0.27	0.35	0.42	0.51	0.34	0.47	0.58
L_1 Error (min)	11.80±14.62*	15.23±19.74***	16.35±19.93***	18.68±21.22***	16.51±20.25***	14.66±19.47***	12.36±18.82	16.89±21.59***	16.00±20.81***	11.49±17.78
Bias (min)	1.03±18.78	(-12.41)±21.64	(-13.49)±21.98	(-14.62)±24.21	(-14.25)±21.91	(-11.00)±21.76	(-6.53)±21.56	(-14.14)±23.49	(-12.32)±23.20	(-5.85)±20.36
SOL_d (min)	25.50±13.84	12.06±12.65	10.98±12.86	9.85±13.45	10.22±11.69	13.47±15.06	17.94±13.80	10.33±12.01	12.15±13.44	18.59±12.76
SOL_{GT} (min)	24.47±23.17									
N	170 recordings									

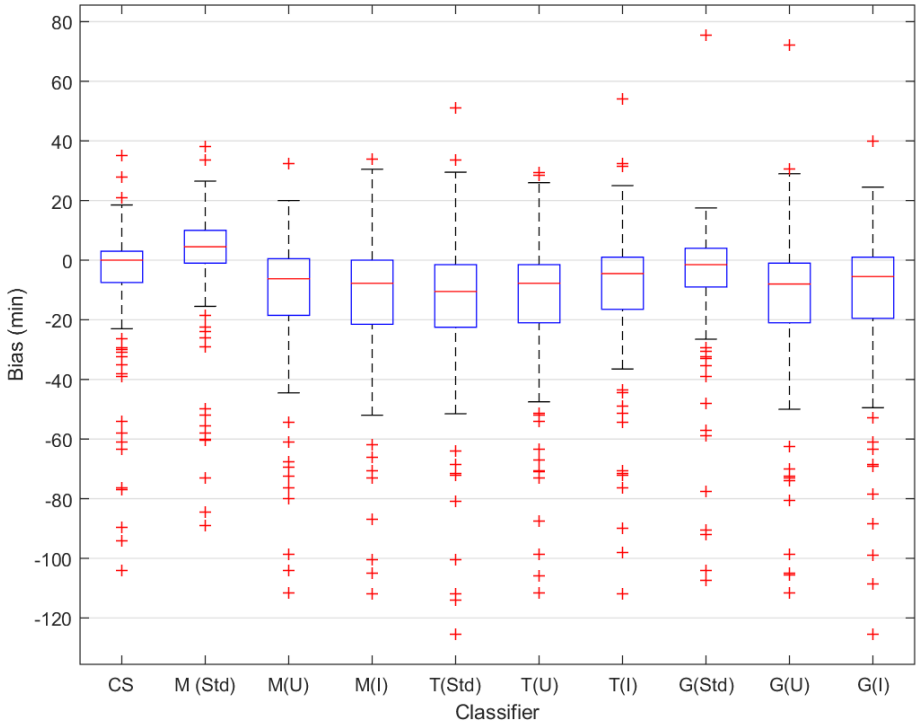


Figure 44.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 110 min.

Table 27.: Performance measures obtained for the first 120 min of recordings, on the testing set, with models trained on the mixed (M), trouble (T) and good (G) groups of subjects regarding the training set. The feature set for FR was altered between: CFS (std) FS; union (U) and intersection (I) sets. The metrics presented include pooled Cohen's kappa for the time considered and mean and standard deviation values on SOL-detection L_1 error, bias, SOL_d and SOL_{GT}. The significance of the differences between the results obtained with new approaches in comparison to current classification (CS) results calculated with a two-tailed Wilcoxon-signed rank test. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Window	**120 min**									
Model	**M (Std)**	**M (U)**	**M (I)**	**T (Std)**	**T (U)**	**T (I)**	**G (Std)**	**G (U)**	**G (I)**	**CS**
κ	0.62	0.31	0.31	0.61	0.32	0.32	0.60	0.32	0.47	0.58
L_1 **Error (min)**	**12.03 ± 14.18***	16.65±20.05***	16.61±19.94***	12.55±15.19**	16.76±20.53***	16.77±20.27***	11.50±18.08	17.26±20.94***	14.77±18.86***	11.49±17.78
Bias (min)	**1.87 ± 18.52**	(-13.99)±22.00	(-14.08)±21.82	0.95±19.70	(-14.80)±22.00	(-14.23)±22.14	(-4.64)±20.94	(-15.32)±22.41	(-9.91)±21.82	(-5.85)±20.36
SOL_d **(min)**	26.34±14.93	10.49±13.18	10.39±12.74	25.42±13.86	9.67±11.44	10.24±12.51	19.84±15.34	9.15±11.59	14.56±13.95	18.59±12.76
SOL_{GT} **(min)**	24.47±23.17									
N	170 recordings									

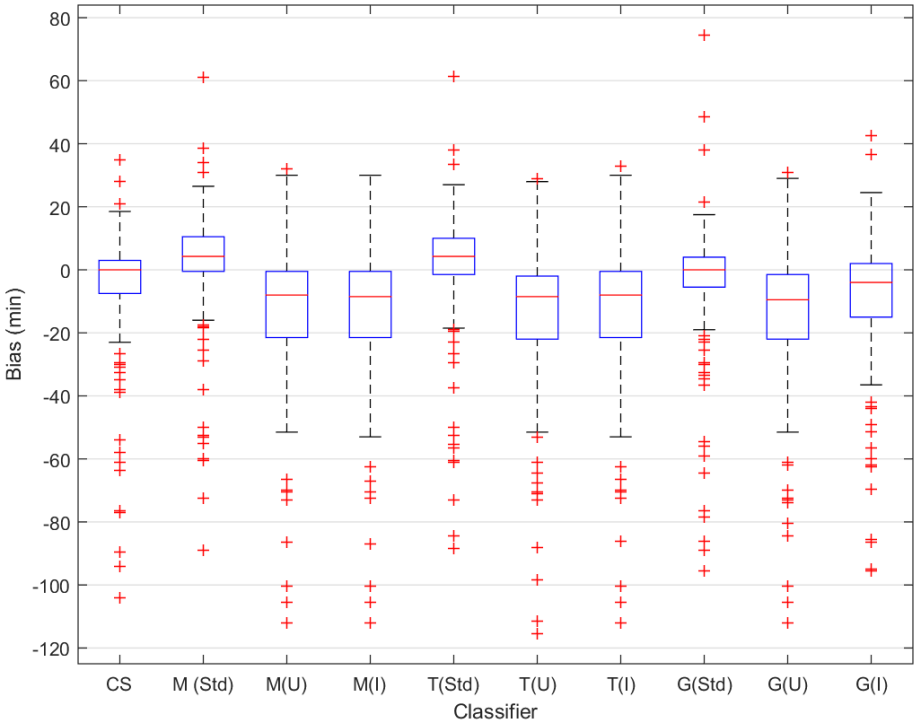


Figure 45.: Box-and-whisker plot, comparing the results on bias estimation of SOL detection, with different classifiers, over a window of 120 min.

A.2 SUMMARY OF THE BEST RESULTS

The box-and-whisker plots on figures 36, 37, 38, 39, 40, 41, 42, 43, 44 and 45 and correspondent tables, show that using a time span under 100 min is not beneficial, when it comes to SOL-detection performance, in comparison to the results provided by current classifiers. On the other hand, with a time window of 100, 110 or 120 min, using certain requirements regarding training population and method of FR, there are significant improvements. Specifically, with a classifier trained on all of the training examples and with CFS FS methods for FR, (classifier M_{std}) and a time window of 120 min, there is a reduction on bias estimation from 5.85 min of underestimation to 1.87 min of overestimation.
Regarding L_1 error, there was a reduction on the standard deviation, which is due to the elimination of the maximum outlier values (associated to underestimation of SOL), even though the maximum upper value found has increased, as shown on table 27. This one outlier, for which an overestimation of SOL of approximately 60 min happened, is highly affecting the mean statistic on L_1 error, of 12.03 min. However, after a two-tailed Wilcoxon-signed rank test, the overall differences between L_1 error obtained with this method and current methods, were found significant at a p-value of 0.05.
For these reasons, we have chosen a window of 120 min, and a classifier trained on all the subjects from the training set, with CFS FS methods for FR, M (Std), to proceed with, forward in this work.

B

FEATURE SETS

Table 28 contains a list of the names of the features (and a pointer for their description in literature, if applied) selected for the purpose of detecting *wake* after SOL. The features set was applied on the training/testing the first classifier of the proposed method for *sleep/wake* detection, described on chapter 6.

Table 28.: List of features used with the second classifier, in the *sleep/wake* classification of subjects with delayed SO_{GT}, with the new proposed method, described on chapter 6.

Name	Literature
Mean of the normalized RR interval (Estimated from the time domain)	[5]
Median spectral power of the ECG [65]	
Continuous wavelet transform over the respiratory signal. Preserves peaks by subtracting the medfilted signal (z-score, medfilt)	[37]
Higuchi measure of phase coordination between respiration and cardiac activity	
Mean of high Teager Energy of RR intervals after EMD (from the first IMF)	
Assortativity mixing coefficient of cardiorespiratory interaction series in visibility graph	[36]
Spectral power in the adaptive High Frequency band (optimized by respiratory frequency)	[66]
Ratio between the adaptive low and high frequency band power	[66]
Scaling exponent of detrended fluctuation analysis for slower time scales	[62]
Samplen entropy of the respiratory effort signal	
Constrained Dynamic Time Warp distance measure (z-score)	[63]
Peak power of the ECG spectral density	[65]
90th percentile of the RR intervals	
Dynamic frequency warp distance measure for resp, over the power spectral density	[63]
Median of the likelihood ratio of heart rate variability	[40]
Number of nodes of interbeat intervals in visibility graph with a small degree (≤ 3)	[36]
Assortativity mixing coefficient of breath-to-breath intervals in visibility graph	[36]
Number of nodes of breath-to-breath intervals in visibility graph with a large degree (≥ 8)	[36]

This work was performed (and funded by) Philips Research (Eindhoven, The Netherlands) under the supervision of Pedro Fonseca and Paulo Novais (Universidade do Minho).