Driver Drowsiness Detection with Peripheral Cardiac Signals

Lourenço Maria Abrunhosa Monteiro Rodrigues

Thesis to obtain the Master of Science Degree in

Biomedical Enginnering

Supervisors: Prof. Ana Luísa Nobre Fred
Prof. André Ribeiro Lourenço

Examination Committee

Chairperson: Prof. João Miguel Raposo Sanches
Supervisor: Prof. Ana Luísa Nobre Fred
Member of the Committee: Prof. Maria Margarida Campos da Silveira

January 2021

Preface

The work presented in this thesis was performed at the company CardioID Technologies (Lisbon, Portugal), during the period February 2020-January 2021, under the supervision of Prof. André Lourenço. The thesis was co-supervised at Instituto Superior Técnico by Prof. Ana Fred.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

This project would not be possible without all the support, motivation and guidance I received.

Firstly, I would like to thank my parents, who have always been a motivation source since my young days of ”wanting to be a scientist”. In everything I am and do today a part reflects you and how you raised and cared for me. To my grandfather, who taught me to question the world as all philosophers do, and to my grandmother, aunts, uncles, cousins and brothers, thank you for all the love and appreciation.

To Afonso, Jonathan, Martim and Rita, who's friendship is a blessing, thank you for all the fun moments, the human troubleshooting and the transformation of problems into challenges.

And to my girlfriend Sofia, an endless fountain of love and appreciation, and the inspiration that allowed the will and perseverance needed to complete this thesis. You're so weird Pew Pew.

I would also like to thank Professor Ana Fred for accepting to be my coordinator, and to Christer Ahlström, who ceded the datatset used to calibrate the IBI corrector and to dimension the models used in drowsiness detection.

Thank you to the CardioID team, where this project was born, and hopefully will continue to develop. Thank you for making me feel part of the team since day one, and for all I have learnt since then. A special note to Pedro Costa, who's amazing work to build the simulator supported the full testing of the system created by this thesis; to Carlos, with whom I have elevated my code organization and to Cátia, David, Eduardo, Roberto for the participation in the experiment, the way their work ended up complementing my own and the ambient they brought to the office.

To professors Arnaldo Abrandes and Pedro Mendes Jorge thank you for volunteering for the driving sessions.

At last, a special thanks to André Lourenço, who's constant guidance plays a fundamental role on the success of this thesis. For the many hours planing, reviewing and suggesting alternatives, as well as the support and confidence in my proposals, thank you.

Abstract

Annually around 1.35 million people die worldwide as a result of road accidents. Of these, 90% occur because of human fault. Such faults have been continuously reduced by the development of safer road architectures and legislation that intends to guarantee the ideal conditions for driving.

However, errors made by human drivers when driving while feeling drowsy result in a constancy of people involved in road accidents, raising the need for a drowsiness detection system. A physiological signal capable of early identifying such state is the heart rate variability, which can be obtained by analysis of the consecutive time intervals between heart beats.

Using peripheral cardiac signals, signals containing cardiac rhythm information and obtained through non-intrusive ways, it is possible to integrate such detection on a vehicle without affecting the driving task.

This work builds the pipeline to use any of three wearable devices: wrist worn PPG band, ECG chest strap and off-the-person ECG collection through a steering wheel, to collect the inter beat intervals, calculate HRV features and detect the drowsiness state of a driver.

A filter was developed to compensate ambient light sensitivity of PPG based devices and the intervals detected from all signals were corrected by an algorithm created to possible wearable contact losses. SVM models with linear kernel and C=0.3 and a selected group of HRV features had good performances , reaching an average 0.62 Matthews correlation coefficient across 12 individuals. Simulator experiments showed good indication that peripheral cardiac signals can be used for drowsiness detection.

Keywords

Heart Rate Variability; Wearable; Drowsiness; Peripheral Cardiac Signals; Machine Learning

Resumo

Anualmente, cerca de 1.35 milhões de pessoas morrem fruto de acidentes rodoviários. Dessas, 90% ocorrem devido a erro humano. Estas falhas têm sido continuamente reduzidas pelo desenvolvimento de estradas mais seguras e legislação que pretende garantir as condições ideais para condução.

No entanto, erros de condutores sonolentos resultam num número constante de pessoas envolvidas em acidentes rodoviários, levantando a necessidade de um sistema de deteção desse estado. Um sinal fisiológico capaz de identificar sonolência é a variabilidade cardíaca (HRV), que pode ser obtida pela análise dos intervalos de tempo entre batimentos cardíacos consecutivos (IBI).

Ao usar sinais cardíacos periféricos, que contêm informação sobre o ritmo cardíaco e são obtidos de forma não intrusiva, é possível integrar tal sistema num veículo sem afetar a tarefa de condução.

Este trabalho constrói o processo para usar qualquer um de três dispositivos: pulseira com sensor PPG, banda de peito e volante capazes de medir o ECG, para obter os IBIs, calcular variáveis de HRV, e detetar a sonolência de condutores.

Foi desenvolvido um filtro especializado para remover artefactos do PPG. Os intervalos recolhidos de todos os dispositivos foram corrigidos por um algoritmo criado para compensar percas de contacto com os sensores. Modelos SVM com kernel linear e C=0.3 e um grupo selecionado de variáveis de HRV mostraram boas performances, atingindo uma média de 0.62 de coeficiente de correlação de Matthews em 12 indivíduos. Experiências em simulador deram bons indícios de que sinais cardíacos periféricos podem ser usados para deteção de sonolência.

Palavras Chave

Variabilidade Cardíaca; Wearable; Sonolência; Sinais Cardíacos Periféricos; Machine Learning

Contents

List of Figures

List of Tables

List of Algorithms

Listings

Acronyms

ADAS	Advanced Driver Assistance System
AECS	Average Eye Closure Speed
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
ApEn	Approximate Entropy
AR	Autoregressive model
BPM	Beats Per Minute
DFA	Detrended Fluctuation Analysis
ECG	Electrocardiography
EEG	Electroencephalogram
ESS	Epworth Sleepiness Scale
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GBT	Gradient Boosting Tree
GSR	Galvanic Skin Response
HF	High Frequency
HR	Heart Rate
HRV	Heart Rate Variability
ICC	Intra Class Correlation

IIR	Infinite Impulse Response
IBI	Inter-Beat Interval
JDS	Johns Drowsiness Scale
KSS	Karolinska Sleepiness Scale
LED	Light Emitting Diode
LF	Low Frequency
MAD	Mean Absolute Deviation
MCC	Matthews Correlation Coefficient
NPC	Non-Player Character
ocSVM	one class SVM
PCB	Printed Circuit Board
PDA	Personal Digital Assistant
PERCLOS	Percentage of Eyelid Closure
PKE	Positive Kinematic Energy
PPG	Photoplethysmography
PSD	Power Spectral Density
PRV	Pulse Rate Variability
PTT	Pulse Transient Time
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Squared Error
RMSSD	Root Mean Square of Successive Differences
SampEn	Sample Entropy
SDNN	Standard Deviation of Normal intervals
SDSD	Standard Deviation of Successive Differences

SENN	Standard Error of Normal intervals
sHRV	short Heart Rate Variability
SMOTE	Synthetic Minority Oversampling TEchnique
SNR	Signal to Noise Ratio
SSS	Stanford Sleepiness Scale
SVM	Support Vector Machine
TINN	Triangular Interpolation of NN interval histogram
t-SNE	t-Distributed Stochastic Neighbor Embedding
ULF	Ultra Low Frequency
usHRV	ultra-short Heart Rate Variability
UUID	Universally Unique Identifier
VLF	Very Low Frequency

1

Introduction

Contents

1.1 Context

Annually around 1.35 million people die worldwide as a result of road accidents [1]. Of these, 90% occur because of human fault [2]. Such faults have been continuously reduced by the development of safer road architectures and legislation that intends to guarantee the ideal conditions for driving.

However, despite all efforts, errors and distractions caused by the insistence on driving even when feeling drowsy result in a constancy of people involved in road accidents.

For this reason, it has become of the uttermost importance to develop systems capable of identifying driver drowsiness, to act with them to prevent in a more personalized and effective way this dangerous behaviour. Several proposed systems are already available in the market, but are usually based on extrinsic factors, as the simple measurement of time driving, or the monitoring of driving behaviour. Even though their implementation on the vehicle is as non-invasive as one could desire, the fact that they monitor only variables external to the driver leads to performances that fall short of what such vital system should.

On the other side, it is known that the monitoring of physiological data allows insight on the internal mechanisms that produce drowsy states, providing an excellent source of information to assess the drowsiness state of any driver. However more powerful information exists in these signals, the technology to read them normally implicates an higher level of intrusion on the drivers environment, which is why they have been kept away from this field of application.

One of the physiological signals that has revealed an interesting capability to identify an individual's drowsiness state is Heart Rate Variability (HRV), which is obtained through the analysis of the series of time intervals that separate heart beats, usually identified through QRS complexes in an Electrocardiography (ECG). Again, the need to place chest electrodes to collect the ECG renders this approach impractical, but, fortunately, less invasive alternatives have been proposed to collect the needed information, measuring cardiac rhythm information in a more peripheral way. These non-invasive technologies combine the feasibility of being installed on a vehicle without disturbing the drivers environment, with the ability to infer their drowsiness state from an intrinsic signal, instead of possible manifestations of such state.

This way this work defines peripheral cardiac signals as the set of physiological signals that measure the cardiac rhythm dynamics in a non-invasive form, that is, which collection doesn't demand any change in drivers routine, or that in any way forces him to have his activity affected by the connection with the measuring devices.

Two pieces of equipment already available seem to meet such criteria, the *CardioWheel*® by CardioID Technologies, a steering wheel cover that measures a bipolar derivation of ECG through the drivers hands, and wristbands and smart watches with an integrated Photoplethysmography (PPG) sensor, as long as these allow continuous signal acquisition, unlike most consumer grade devices that only report an average value of measure heart rate.

However, while both of them exceed expectations when it comes to practicability, fitting perfectly into anyone's lifestyle and driving, the distancing from the cardiac signals' primary source demands a more careful processing of these peripheral signals in order to extract information as trustworthy as that collected with thoracic electrodes.

Having this, this thesis proposes to answer two main questions, that ultimately combine to produce a drowsiness detection system for driving environments based on peripheral cardiac signals: How to deal with the processing of such signals (filtering and fiducial points localization), and ensure confidence on the HRV features obtained from them, and if the HRV information obtained from these sources allows such drowsiness state classification as it does with thoracic ECG.

1.2 Objectives

The two questions proposed before define the main objectives of this theses, which should be further specified briefly here.

1.2.1 How to process peripheral signals?

Processing strategies differ depending on which equipment is used, while the *CardioWheel*® has filters built in, and so, returns a signal where the QRS complexes are immediately identifiable, most wrist band PPG sensors return a raw signal, filled with noise, movement and light change artifacts. For this reason an online filtering strategy is implemented to extract only the pulsated component of PPG.

Having the peaks on these signals corresponding to ventricular systole (ECG) and systolic pulse (PPG) peak detection algorithms are implemented to store the timestamps at which heart beats occurred. This stream of timestamps is then used to calculate the series of Inter-Beat Intervals (IBIs) (fig. 1.1) that are the base of HRV calculation.

As the continuity of the peripheral cardiac signals depends on the constant contact of hands on wheel, or the absence of too strong artifacts on the PPG, it cannot be guaranteed at all moments. Thus it is possible to have missed peaks or false detections in the stream of heartbeat timestamps. To solve this, an IBI correcting algorithm is proposed, and validated, artificially removing or adding peaks and evaluating the error remaining after correction.

It is to note that the wearable nature of these signal's source results in that continuous segments may have duration a lot shorter than what state-of-the-art HRV use, even with correction. Having in mind that short Heart Rate Variability (sHRV) need a minimum 5 minutes of uninterrupted inter-beat intervals detected, and, in wearables, 2-3 minutes would be a much more realistic projection, the analysis

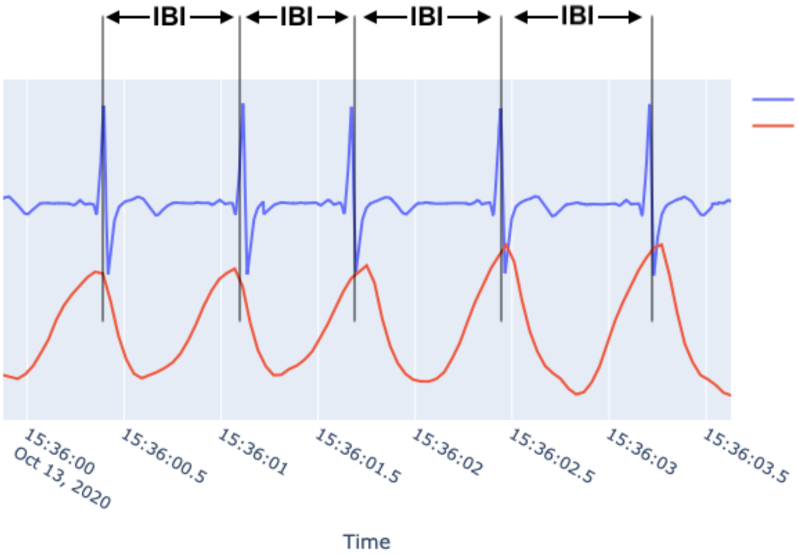


Figure 1.1: Representation of Inter-Beat Intervals on ECG and PPG signals. IBIs are the interval of time that separates two consecutive heart beats.

conducted must be redirected to the field of ultra-short Heart Rate Variability (usHRV), and, to do so, the validity of features in this ultra short scope will be assessed before using them in classification models.

1.2.2 How to classify drowsiness from peripheral signals derived HRV?

To start building a classification model on this subject, an already existing dataset containing naturalistic driving data, with both ECG measurement and drowsiness annotations is used to evaluate machine learning algorithms in their capacity to correctly output drowsiness alarms from usHRV features. This database is also used to evaluate the need for class balancing, feature selection and alternative training strategies.

After defining the optimal models and training procedure, data collected with a driving simulator (fig. 1.2) developed by CardioID/ISEL is used to evaluate the performance of such models when using peripheral signals as data source. This dataset contains drowsiness annotations, hands ECG from *CardioWheel*®, wrist PPG measured with the pulseOn wrist band, and chest ECG from a Movesense® chest-band. Positive results in this section establish a system using peripheral cardiac signals only to detect drowsiness in drivers, combining the non-invasive advantages of wearables and built-in vehicle systems, with the deep insight physiological signals provide.

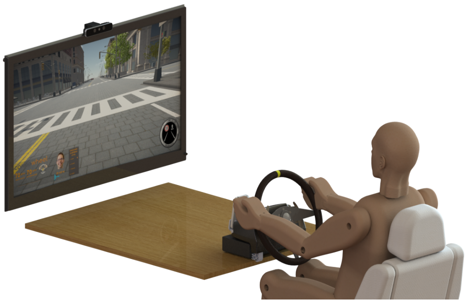


Figure 1.2: Driving simulator setup: A computer simulated environment is presented in the screen, while the driver controls a vehicle using the pedals and Cardiowheel. The simulator not only integrates the inputs to run the environment, but also aggregates inputs from the wheel movements, CardioWheel sensor, and intel realssense camera to a database.

1.3 Thesis outline

In order to fulfill the defined objectives, a comprehensive literature review is performed in chapter 2, where not only currently used tools in this field are presented, but also a theoretical basis is used to justify the relations here proposed. In chapter 3, the state-of-the-art of cognitive state characterization in drivers is , namelly stress and, more importantly, drowsiness. Chapter 4 explicit the methods implemented to filter the peripheral signals, as well as to correct the IBI stream they provide. A report on drowsiness classification from HRV features is present in chapter 5. Finally, an experiment is setup to test the system developed for signal extraction and drowsiness classification and described in chapter 6.

2

Literature Review

Contents

2.1 Photoplethysmography

Photoplethysmography (PPG) was first described by Alrick Hertzman in 1937. He noticed that the amount of infrared light absorbed by the tissues varied along time, so it was proposed that this change depended on the volume of blood passing the tissue bed at each time. This founds support in Beer-Lambert's law, where light absorption is essentially dependent on the path distance the light rays travel, concentration and the specific absorption coefficients of each substance for different wavelengths.

In fact, with every heartbeat, a pulse of new arterial blood passes the arterioles in locations as the wrist, finger or earlobes. As this increases the amount of light absorbed, a high correlation between PPG pulse intervals and heart rate can be found and, from this, Pulse Rate Variability (PRV) can be used as an estimate of standard HRV [3].

An example of a typical PPG waveform can be seen in Figure 2.1. As stated before, this waveform describes the variation in blood volume through time. As it is possible to observe, a fast increase in volume results in the systolic peak, which corresponds to the pulse of blood incoming from the heart after systolic contraction. From this point, blood volume should decrease steadily, however, as this signal is also sensitive to the reflected pulse, *i.e.* blood returning to the heart, a second, smaller peak, the diastolic peak, can also be detected. This specific format gives this signal a richness of information to be extracted from it and its derivatives. While most of these information is used currently to estimate, not only heart rate, but very different physiological aspects, as cardiac output volume, arterial pressure and stiffness, and even blood oximetry [4], the interest of this work is to extract the precise time intervals between pulse peaks (IBI) and calculate HRV features that allow drowsiness detection.

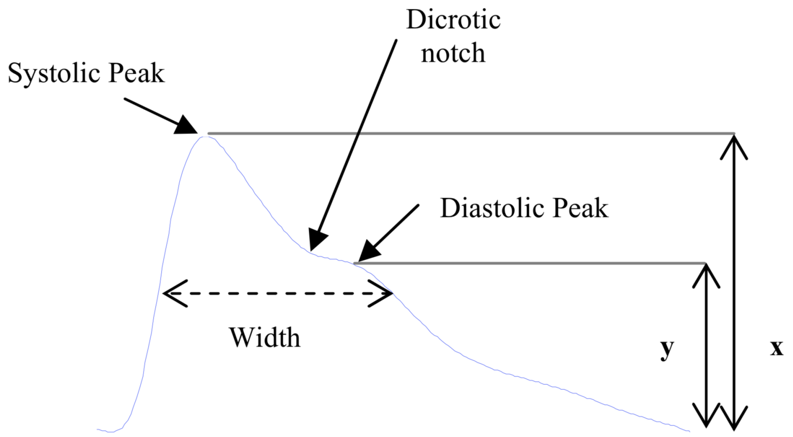


Figure 2.1: Schematic of typical PPG pulse waveform, with most relevant features and key points denoted.

2.1.1 Motion Artifact Filtering

Being an optic method, PPG is extremely sensitive to phenomena that result in a change of detected light intensity or hemodynamics. These include motion of the measuring site or sudden alterations in the ambient lightning, which will result in the superposition of these perceived changes on the real PPG signal. Also, given that these light variations have a power much higher than that of the subtle blood light absorption, they will completely hide the information needed and must be handled before any analysis on the signal is performed. As most of the abrupt lightning changes correspond to discontinuities, and other artifacts produced by motion and respiration have frequency ranges bellow the PPG pulse, filtering the signal can mitigate their presence in the signal.

It is to note that, as stated by Park, Waugh and his colleagues [5, 6], phase distortions can occur if the phase response of the applied filter is not taken into consideration, which could ultimately result in completely invalid results after HRV analysis of the filtered signals. Because of this, Finite Impulse Response (FIR) filters with linear phase response are used and, when Infinite Impulse Response (IIR) filters are needed for their lower orders, a forward-backward design has to be implemented. However, the latter does not allow real-time applications, which, in the scope of drowsiness detection during driving tasks, narrows the filter choices to linear phase FIR.

A table (2.1) containing proposed filtering methods for PPG is presented bellow, which will serve as a justification for the methods implemented in this thesis.

Source	Year	Filter Type	Phase distortion correction
Sabeti *et al.* [7]	2019	Butterworth Band-Pass Filter	None
Liang *et al.* [8]	2018	4th order Chebychev Low-Pass	Forward-backward design
Waugh *et al.* [6]	2018	FIR Low-Pass filter IIR High-Pass filter	Symmetric FIR Phase non-linearity at very low frequencies
Park [5]	2017	Harmonic IIR Notch Filter	Forward-backward design
Subhagya *et al.* [9]	2017	Band Pass-filter LMS-adaptive filter	None
Ye *et al.* [10]	2017	Adaptive Noise Cancellation Singular Spectrum Analysis	None
Sun [11]	2012	Moving Average	Symmetric FIR
Ram *et al.* [12]	2012	Adaptive step-size Least Mean Square adaptive filter	None
Wei *et al.* [13]	2011	Median Filter FIR Low-Pass Filter Wavelet decomposition	Symmetric FIR

Table 2.1: Filters used to preprocess PPG signal in literature.

It is important to note that the works that present none as a method for phase distortion correction simply didn't mention it, leaving the possibility that symmetric FIR or forward-backward designs were used but not made explicit in those papers. Also, as some of this works [7, 9, 12] were not concerned with a precise characterization of time features, as peak position, but only in morphological classification of pulses or even oxygen saturation estimates, phase distortions could have been disregarded as they would not affect those results considerably.

Given the difficulty that superposition between PPG and artifacts presents, some researchers coupled their filtering processes with automatic classification of pulses and signal quality assessment so that non-correctable segments of it are identified as such and discarded before any signal parameters are derived.

A summary of the methods and classifiers used in this signal classification are depicted in Table 2.2.

Source	Year	Features	Classifiers
Sabeti *et al.* [7]	2019	Amplitude Pulse duration Peak-Peak jump Valley-Valley jump	Decision Tree Ensemble Decision Tree SVM Threshold Optimization
Waugh *et al.* [6]	2017	Normalized Pulse against template	Clustering: Pearson Correlation Kendal Rank Correlation Spearman Rank Correlation RMSE
Karlen [14]	2012	Amplitude Maximum & Minimum intensity Pulse period Slopes Distribution	Adaptive Thresholding
Sun [11]	2012	Amplitude Dissimilarity with template Onset location Pulse peaks location and number Pulse duration Peak-Peak interval	Thresholding Kalman filter

Table 2.2: PPG pulse classifiers in literature.

All of these classifiers are based on a definition of what is a ”normal” PPG pulse, and serve as an excellent pointer of which features of this waveform provide the most accurate description of it, features as amplitude and pulse duration are transversal for all methods, and some other interesting characteristics can be investigated, as a quantification of all peaks present in a pulse, and the characterization of the distribution of slopes in it.

2.1.2 Peak detection

Being a tool for Inter-Beat Interval determination, it is crucial to be able to accurately detect the peaks correspondent to the systolic beats. Differently from ECG where QRS complex has a very distinctive morphology when compared with the rest of the signal, and is very narrowly time localized, systolic beats in PPG can be a little bit more complicated to identify if some noise or even if an abnormally large diastolic peak is present, potentially compromising the temporal resolution of its peak detection [15].

For this, different techniques to find the correct peaks in PPG signal have been proposed, which take advantage of the derivatives of this signal or some other heuristics based on the relation it has with physiological behavior of the heart, as refractory periods[1] (Table 2.3).

Source	**Year**	**Peak detection criteria**
Vadrevu [17]	2019	Zero-frequency resonator correlation
Thang [18]	2017	Adaptive threshold Refractory period Amplitude variation
Wei *et al.* [13]	2011	Derivative zero-crossing
Shin [19]	2009	Adaptive threshold Refractory period

Table 2.3: PPG peak trackers in literature.

2.1.3 Peak location refinement

The digitization of PPG measurements places a dependency of peak detection precision on the system sampling frequency. Even though very high frequencies, around 1000Hz, are technically possible and even used in clinical settings, wearable form factors require lower sampling to optimize power consumption and battery autonomy. This results in a tendency for reduction of sampling frequency in most wearables, reaching frequencies as low as 25Hz.

The investigation on the effects of this sampling level have been investigated by Choi and Shin [20], who decimated high frequency PPG recordings and compared the accuracy of Pulse Rate Variability at frequencies raging from 5kHz to 5Hz. By comparing the deviation of PRV parameters, obtained from stable PPG signals at different sampling frequencies, and the HRV corresponding ones, from ECG at 10kHz, a measure of how much this downsampling affected the reliability of the obtained data. They concluded, as expected, that the larger the decimation, the larger the deterioration of the parameters reliability, and significant differences were found for parameters as NN50 and pNN50 when sampling frequency is bellow 25Hz. Other parameters are also affected, being observable that time-domain parameters are more sensitive than frequency-domain ones. The authors also underline that their findings were based on very stable PPG, so that systems where the signals are collected in real world applications should take into consideration the additional uncertainty that the inevitable noise brings, and be careful with sampling frequencies too close to this proposed limit of 25Hz.

Additional work has been done in order to investigate the feasibility of using interpolation techniques to regain the temporal resolution lost because of low sampling rates, which has been summarized by Berés [21]. Not only this work presents the results from previous studies, it unifies those findings in a

[1] The fact that heart muscle cells are not immediately ready to contract again after a heart beat [16]

comprehensive manner by generating an artificial PPG correlated signal at 1kHz, decimating it to frequencies between 2 and 500Hz, and cubic spline interpolating it back to the original sampling frequency. After that, HRV parameters were estimated and compared with the original signal, establishing this way a measure of interpolation usefulness in recovering the information lost due to undersampling. In general, they found interpolation to greatly improve the HRV study accuracy, determining nonlinear parameters as Poincaré-plots to be the most sensitive ones to this process, which required minimum 10Hz original sampling frequency so that the interpolation could reconstruct the signal. However, we must note again that the signals used were artificial, presenting no artifacts and a controlled variability. Because of this, the uncertainty added by such factors in real signals must be taken into account, and this minimum limits should be avoided, in order to guarantee an error margin and reliable results at the end of the analysis.

Finally, two different interpolation techniques are used in literature to refine peak location estimates, cubic spline interpolation and parabola approximation.

Cubic spline interpolation defines a C^2 function that is a cubic polynomial between each pair of consecutive sampled points. Even though it has had its usefulness tested for PPG peak refinement, it is computationally expensive as it needs to solve a set of three equations for each pair of points, relating adjacent segments' derivatives. This can be avoided using a simpler method, the parabola approximation. This is a method used only for peak estimation, domain at which it maintains an accuracy close to that of cubic spline, but, as it uses a quadratic approximation of the observed set of points, it is able to rewrite the needed calculus into just one single equation. I does so by defining a three sample set, where the central one is the observed local *maxima*. Using this points, and rewriting the parabolic equation allows the direct definition of peak location both in time and amplitude directly from the sample values. Baek *et al.* [22] compared the performance of both cubic spline and parabola approximation in PPG peak estimation and found their results to be very similar, while the computational burden of the second was consistently lower.

2.2 Heart Rate Variability

HRV is the study of variation of consecutive Inter-Beat Interval over time. This type of observation provides a window to perceive the balance between systems responsible for the modulation of cardiac rhythm, namely the Autonomic Nervous System (ANS) sympathetic and parasympathetic systems, and the identification of anomalous intervals that can be correlated with cardiac disease. Because these clues may not happen at all times, but only in specific periods of the day, the need for continuous monitoring of heart rate appeared, carrying with it the time consuming task of analysing the accumulated data. This resulted in the appearance of computational methods to form indexes that would condensate all the observed data and point out if some worrying information is present. [23]

2.2.1 HRV and ANS

For this work it is specially important to establish a relation between HRV and the balance between sympathetic and parasympathetic systems, and to understand how different psychological states influence that balance.

As stated in [24], the ANS is constituted by two antagonist systems: the sympathetic, that in a general form prepares the organism for energy expenditure and stress response, and the parasympathetic, that returns the body to its basal, relaxed state.

Sympathetic system, also referred to as the ”fight or flight” system, produces a series of alterations in the body, such as vasodilatation of coronary arteries and vasoconstriction of other vessels, as well as increase in heart rate. This optimizes cardiac output and oxygenation of muscles, that must be optimally active to respond to the stress source.

Contrarily, parasympathetic system will promote a restful state, dilating peripheral circulation and slowing down the heart rate, so that other systemic functions, such as digestion and lachrymal, saliva, urine and fecal secretion take place. For that reason it is called as the ”rest and digest” system.

In healthy subjects, both branches of ANS balance each other, with sympathetic predominance representing an active acceleration of Heart Rate (HR) and the parasympathetic dominance a passive return to the basal state, with consequent deceleration of HR. These effects are observable in HRV studies, specially through frequency domain parameters, as their continuous balancing process produces oscillations at defined frequency ranges.

It is widely accepted in the scientific community that two different bands of spectral analysis of HRV correlate with distinct activity levels of ANS branches [23, 25, 26]. Specifically, high frequencies (0.15 to 0.40 Hz) are commonly related to parasympathetic activity, while low frequencies (0.04 - 0.15 Hz) can be related to a mixture of both activities, or, as some researchers propose [26], sympathetic activity alone if frequency band powers are in normalized units. This allows the direct evaluation of ANS state through HRV as a non invasive form of accessing a person’s internal state.

Other HRV indexes provide additional information, and can be divided in three main groups, depending on the type of analysis needed to calculate them: Time domain, frequency domain and nonlinear domain. These are presented in the following sections.

2.2.2 Time Domain

Time domain parameters result from the direct analysis of inter-beat intervals present in the tachogram (fig. 2.2), again, two subgroups are described [23]: statistical indexes and geometric ones. Both are directly calculated from the series of IBI values, and several statistical parameters have been described, such as Standard Deviation of Normal intervals (SDNN), Standard Error of Normal intervals (SENN), which is the standard deviation of the sample distribution means, Standard Deviation of Successive Differences (SDSD), which is the standard deviation of adjacent intervals' difference, Root Mean Square of Successive Differences (RMSSD) and the number of adjacent intervals that differ more than 50ms (NN50) and the relative homologous, pNN50, that is the percentage of NN50s in the entire sample. Geometric indexes are built upon a geometric representation of the RR intervals, as a histogram of those intervals. Two indexes are presented in the review by Acharya *et al.* [23] the triangular index, that results from the quotient between the total of intervals analysed and the amount of intervals that fall in the modal duration (peak of the histogram), that can be viewed as the division between the are and the height of a triangle approximating the RR interval distribution [26], and the Triangular Interpolation of NN interval histogram (TINN), that measures the width of the triangular approximation of the RR intervals histogram as a measure of HR variability. Geometric analysis provides on clear advantage, that is is insensitivity to data quality, as the triangular approximation acts as a automatic outlier removal process. However, these methods need at least 20min of recorded data [26], which renders them unsuitable for short and ultra-short term HRV studies.

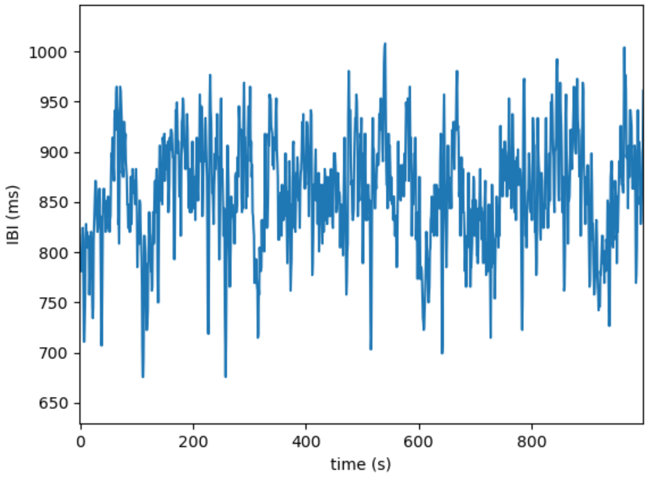


Figure 2.2: Example of tachogram, it represents the time series generated by consecutive inter-beat intervals, and it is the basis of HRV analysis.

2.2.3 Poincaré Plot

Poincaré plots are a representation of detected RR intervals as a function of the previous interval (fig. 2.3, as is commonly seen in state space diagrams from control theory, which distribution and formed geometry can afterwards be used to access the normality of a given tachogram. Two simple indexes are calculated from this plot, that correspond to the variance along the directions $NN_n = NN_{n-1}$, successive intervals maintain the same duration (SD_1), and $NN_n = 2NN_{mean} - NN_{n-1}$, successive intervals vary proportionately to the distance to the mean interval duration (SD_2). This gives a measure of the different contributions of long term and short term variability respectively. The ratio SD_1/SD_2 can also be obtained to perceive the balance between these two components [23].

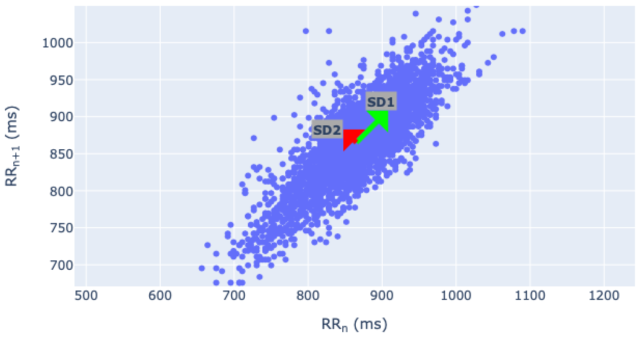


Figure 2.3: Example of Pointcaré plot, with SD1 and SD2 features depicted.

2.2.4 Frequency Domain

Frequency domain parameters are found through the measurement of variance or power in defined frequency bands of the Power Spectral Density (PSD) representation of the collected RR intervals. Two different approaches can be used to estimate PSD, a non-parametric method as Fast Fourier Transform (FFT), or the parametric Autoregressive model (AR). While FFT is a simple and fast method that produces a discrete frequency decomposition of the original signal, the AR is more complex and need verification of its model suitability [26] but results in a continuous and more accurate PSD estimation. Four frequency domain indexes are commonly used, that are the powers measured in defined frequency bands: Ultra Low Frequency (ULF) for frequencies bellow 0.003Hz, Very Low Frequency (VLF) in frequencies between 0.003Hz and 0.04Hz, Low Frequency (LF) for frequencies that lie in a 0.04 to 0.15Hz interval and High Frequency (HF) for those up to 0.4Hz [23, 26]. It is yet important to note that these bands are not resolvable independently of the record duration, this because some of the lower frequency bands measure oscillations that take longer periods of time than the recordings themselves. As an example, to measure phenomena bellow 0.04Hz oscillations longer than 25 seconds are need and, for 0.003Hz, 5min30s are needed. Because of this, even with continuous spectrum from AR, measurements shorter than 5 minutes cannot provide any reliable information about ULF [26], and even the interpretation of VLF powers must be done very carefully in this time frames, as the proximity to these frequency resolution frontiers becomes very relevant.

2.2.5 Nonlinear Domain

Nonlinear or fractal parameters are derived from chaos theory, which accounts for non-linearity and non-periodicity of signals in systems as physiological ones [26]. There is also evidence that this approach to evaluate HRV is more accurate than the previously defined indexes [23, 26] as it is able to accept that heart rate is slightly random and sometimes chaotic.

Godoy [27], presents a comprehensive review on nonlinear methods for HRV, dividing them into tow categories invariant and information domains. Invariant domain contains parameters like Fractal and Correlation Dimension, Detrended Fluctuation Analysis, Hurst exponent and Largest Lyapunov Exponent, that give measures of signal regularity, that is, larger values express more complex variations of HR and lower ones a more regular rate, being the first associated with healthy subjects capable of quickly adapting their cardiac system to other *stimuli*. Information domain comprehends for methods of entropy analysis, Approximate, Sample, Shannon and Multiescalar entropy. These reflect the randomness or unpredictability of a signal, so that higher entropy levels refer to more irregular HR.

While nonlinear HRV indexes seem to better measure the irregularity of HR, as it deals directly with chaos and randomness of signals, it is also very sensitive to low quantities of information, as stated by Acharya *et al.* [23], so one must be careful employing nonlinear analysis in shorter tachograms, as they may not contain enough information to separate chaos from noise. This sensitivity renders most non-linear indexes unusable for usHRV, as with 2 minutes or less of signal, around 120, at most, IBIs are available for analysis.

2.3 Pulse Rate Variability

HRV is usually made using ECG recorded tachograms. However, this is expensive and uncomfortable for most measurement scenarios, as the most portable solutions for ECG measurement consist of band straps around the chest. For this reason, PPG based devices present a promising alternative, as they can be presented in very practical form factors as smart bands and watches. Nonetheless, this alternative has to consider that the physiological mechanisms measured by these systems are not the same, ECG measures the electrical activation of cardionector tissue in the heart while PPG the volume change in peripheral circulation that blood pulses provoke. This means that differences between the two of them are unavoidable, the ECG can instantaneously detect ventricular systole, but PPG will only find its peak after a delay, that corresponds to the time the blood takes to travel from the heart to the measurement place. This time is referred to as Pulse Transient Time (PTT), and has several variables influencing it, thus, influencing the precision of PRV too.

A study conducted by Murakami and Yoshioka [28], related HR and PTT. Though our goal is not to verify this relation, it is to justify that the variation in PTT does not compromise the correlation between

HRV and PRV. In fact, though PTT seems to vary linearly with heart rate, the standard deviation of PTT when the standard deviation of IBI was around 50ms was just 4ms, so we can assume that variations in PTT will only cause minor differences between PRV and HRV.

A large set of articles was produced in order to relate these two analysis, by directly comparing both of them.

In 2006, Bolanos [29] built a Personal Digital Assistant (PDA) system to capture simultaneously PPG and ECG signals. From these, time domain parameters (minNN, maxNN, mean, mode and SDNN), frequency ones (Total power, HF, LF and their ratio) and statistical measures of the heart rate signal (variance, skewness, kurtosis and Approximate Entropy (ApEn)) were compared using correlation metrics. From the recordings taken from two subjects (one female), each with three different recordings, very high correlations between simultaneous PRV and HRV was achieved, above 0.99. This opened the door for additional studies that could base their research on this positive result, and try to refine the level of confidence with which PRV constitutes a surrogate for HRV, and in which conditions that substitution is feasible.

In 2015 Jeyhani [15] confirmed these findings using a larger study population, with 18 healthy male individuals. By measuring the relative error between PPG derived parameters and those derived from ECG it was found that most indexes had errors bellow 6% (SDNN, RMSSD, SD1 and SD2), while pNN50 showed to be more sensitive with errors around 29,89%. This study focused also on which fiducial point of PPG produced the better PRV indexes, testing the pulse peak and its second derivative maximum, and, in their trial, the signal peak gave the best results.

Pinheiro *et al.* [30] investigated this with a more robust approach, using a population of 33 healthy subjects and other 35 that suffered from some type of cardiovascular disease. They compared the statistical agreement of SDNN, SDSD, RMSSD, NN50, pNN50, VLF, LF, HF and low-high frequency ratio obtained from ECG and PPG simultaneous recordings. To derive the PRV parameters, an extense list of fiducial points was tested, namely:

- PPG_{onset} - minimum at valleys
- $PPG_{20\%}$ - point at 20% of pulse amplitude
- PPG_{der} - point corresponding to the first derivative maximum
- $PPG_{50\%}$ - point at 50% of pulse amplitude
- $PPG_{80\%}$ - point at 80% of pulse amplitude
- PPG_{peak} - maximum at peak of pulse

Finally three different test groups were created, healthy subjects at rest, healthy subjects after exercise and cardiovascular patients at rest. With this they concluded that PPG is a good surrogate of

ECG for healthy resting individuals, specially if using the derivative peak as a fiducial point for inter-beat interval measurement, and that for the other two groups only some parameters maintain a good correlation with ECG derived HRV, the frequency domain ones, mean and SDNN, specially if $PPG_{20\%}$ and PPG_{onset} are used respectively. It seems that there is not a single fiducial point that proves to be optimal for every situation, so that it should be adjusted for the specific population and scenario where PRV is conducted. Also, exercise and poor cardiovascular condition seem to reduce the proximity between these two branches of cardiac rate analysis. This happens because PRV is not only affected by direct changes in sinus rhythm, but also by vascular stiffness, different intra-thoracic pressure, etc.

More recently, in 2018, Vescio [31] compared PPG and ECG derived inter-beat intervals and HRV parameters using short-term analysis (5 minutes long recordings), with PPG being recorded with a earlobe wearable form factor. Cross correlation and Root Mean Squared Error (RMSE) were used to quantify the similarity between differently obtained indexes. Very little differences were found comparing the inter-beat interval sequences, with a RMSE around 4.4ms. Also, no significant differences between HRV parameters from 1440 segments (from 10 different subjects) were found, except for the mean NN. However, this adverse result is explained by the sensitivity of pair tests to mean shifts, as in fact this difference had a very low dispersion. Poincaré plots supported the measured similarity level between HRV and PRV.

2.4 Ultra-short HRV

After a review by Georgiou [32] (2018), where the accuracy of HRV obtained from signals collected by different wearables (16 ECG chest bands and 2 finger probes PPG) is compared with clinical grade holter ECG, it became clear that increasing movement resulted in the deterioration of this correlation. This happens because of the increased noise and artifact presence, that make some of the heart beats undetectable and some false detections inserted. This results in fragmented signal, and so, shorter continuous segments from which to estimate HRV parameters. The fact that device used to monitor heart rate places a time cap in our analysis, that can very easily be lower than 5 minutes, makes it mandatory to consider a different scope of HRV that fits the small time-windows offered, usHRV.

In 2014, trying to reduce the time constraints short term HRV guidelines demanded (5 min of supine resting and 5 min recording) [33] to potentiate the implementation of this tool in train planing of college athletes, Esco and Flatt [34] studied the agreement of a commonly used in sports HRV parameter, lnRMSSD, when measured in shorter time frames (60, 30, 10 seconds) instead of the normally required 5 minutes. Statistically significant differences were only found in post exercise lnRMSSD for 60 and 30 second measurements, however, those differences were considered trivial after Cohen's d was used to determine effect size. All parameters presented near perfect ($>$0.89) or very large ($>$0.79) Intra Class

Correlation (ICC). Bland-Altman plots were used to test the limits of agreement, and it was concluded that the shorter the segment, the larger these limits would become. This shows that it is possible to shorten the duration of HRV and still obtain relevant results, but that that shortening is not without a cost, and a researcher must first be certain that the usHRV parameters he is using maintain their significance under the measuring conditions and duration used in its application.

Castaldo, Melillo and Pechia conducted a series of studies focusing on the ability of different HRV parameters to maintain their significance in ultra short scenarios, and their feasibility as stress indicators. In 2015 [35], a list of 10 parameters that were significantly linked to stress detection in short HRV was tested: mean NN, LF, LF/HF ratio, Sample Entropy (SampEn), correlation dimension, Detrended Fluctuation Analysis (DFA) long and short term slopes, recurrence plot mean line length and recurrence rate. The Stroop Color Word Test was used to induce stress in 42 participants, who had 7 minutes of ECG recorded before and after the test. The described parameters were calculated from those recordings using the last 5 minutes (short) and 2 minutes (ultra short) from each, and compared to see if significance was maintained and if the variation in both domains was coherent when the analysis is shortened. Even though all of them were coherent, only six kept significance when usHRV was used: mean NN, LF, SampEn, DFA long and short term slopes, recurrence plot mean line length. with this, they concluded that these 6 indexes are useful to detect stress in usHRV of 2 min. However, the lack of significance of the other four should not be interpreted as a definitive discard of them, but as a need to perform studies in larger populations to confirm if they are really not significant.

Later, in 2018 [36], they made a review on existing work on usHRV and identified the lack of a rigorous method to assess ultra short indexes reliability. Most works would use only statistical tests, and some of them wrongly, as assuming that $p\text{-}value \geqslant 0.05$ means the confirmation of null hypothesis, and others would find correlation values without statistically validate them. To provide future investigation on this subject with a standardized solid method for ultra-short/short term HRV indexes surrogate relation, the algorithm presented in Figure 2.4 results from this review.

The authors also note that the normality of the data should also be tested prior to the application of this validation method, so that suitable methods are selected. For example, if the data is not normally distributed, non-parametric correlation methods have to be used instead of parametric ones, the Bland-Altman plot has to be adapted to accommodate this and instead of Cohen's d coefficient, a Cliff's delta statistic should be used to test effect size.

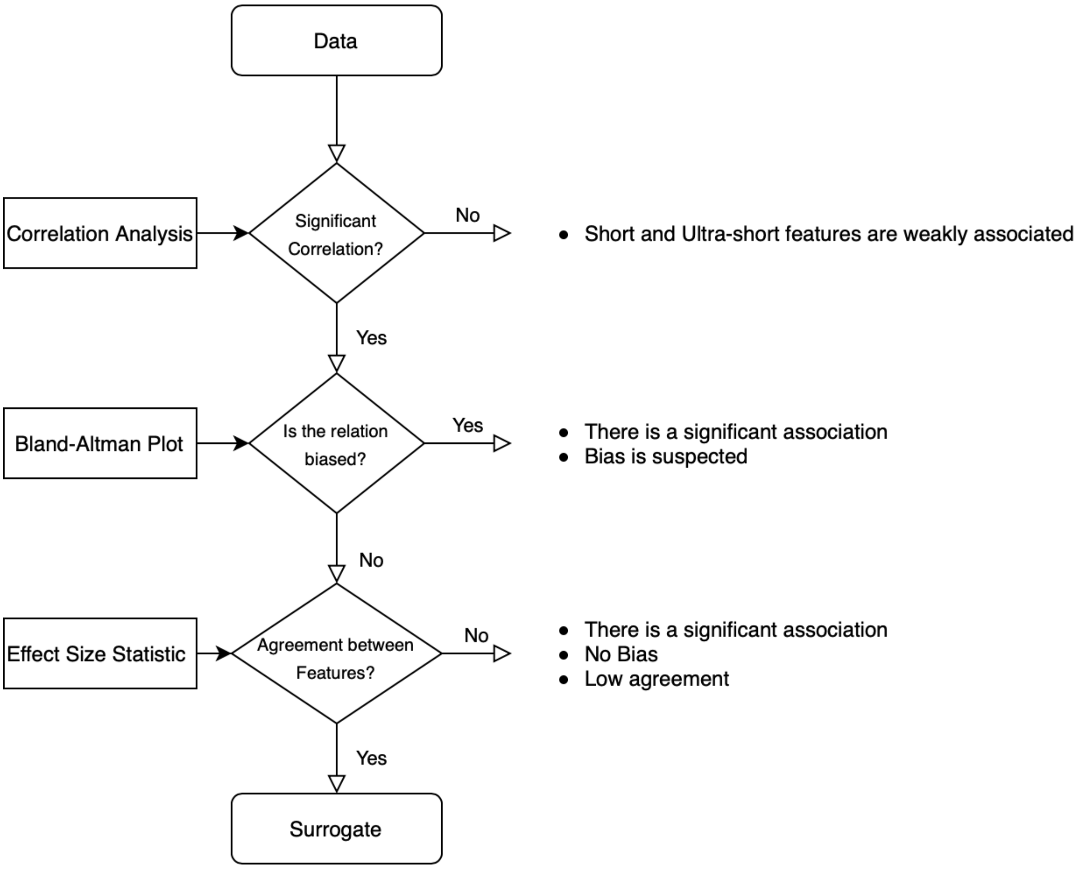


Figure 2.4: Proposed method for ultra short surrogate of short HRV indexes validation, Pecchia *et. al* (2018)

To study the feasibility of using ultra short surrogates for HRV, a new research on usHRV features for stress detection was conducted [37], following the method previously described. 42 healthy subjects had 5 minute long ECGs recorded in two distinct moments, one during an oral examination used as a stressor, and another one after spring break in a calm and relaxed environment. From the collected 5 minute excerpts, shorter ones of 3, 2, 1 and 0.5 minutes were extracted, and they constituted the set of ultra-short segments to be compared with the short (5 min) ones. A total of 23 HRV parameters were tested, which are listed as follows:

- **Time Domain:** Mean NN, SDNN, Mean HR, Standard deviation of HR, RMSSD, NN50 and pNN50
- **Frequency Domain:** LF, HF, LF/HF ratio and Total Power
- **Non-linear Analysis:** Poincaré plot SD1 and SD2 indexes, Approximation, Sample and Shannon Entropy, Correlation Dimension D2, DFA long and short term indexes, Recurrence Plot Analysis mean and maximum line length, and determinism index and the Recurrence Rate

The parameters were computed for every segment, with the exception of some parameters that, according to the guidelines by the task force of the european society of cardiology and the north american society of pacing and electrophysiology (1996) [33], can not be calculated using such short recording periods, such as LF that needs at least two minutes and HF that requires 1 minute of tachogram to be validly obtained. First, to test the feasibility of using each usHRV parameters as a stress indicator, the comparison between medians of each one of the parameters derived from at rest segments and the corresponding indexes at stress. This relation would hold if the trend of means was consistent for different time scales comparatively to the 5 min experiment. Also, validity of this was asserted with Wilcoxon's test $p\text{-}value < 0.05$. Afterwards, even though potential ultra-short term stress indicators are selected, their validity as surrogates of short term ones has to be proven. This was done by following the method described in [36]: Correlation analysis on all time scales, usage of Bland-Altman plots to discard the existence of bias in the found correlation. This produced a set of six features that are consistent surrogates of short term HRV indexes in ultra short analysis: Mean NN, SDNN, Mean HR, Standard deviation of HR, HF and SD2. Moreover, the authors tested these in an automatic classifier scenario to both find an optimal subset of features that eliminated redundancy, and to evaluate their performance in stress detection. Mean NN, Standard deviation of HR and HF formed this subset that achieved over 88% accuracy when employed in an automatic classifier.

2.5 Sleepiness scales

Having seen how to collect and process cardiac dynamics into HRV features, to identify sets of values as the states they correspond too, some quantification method needs to be used. In this work we will focus on the detection of alert vs. drowsy states, so a review on different methods to measure sleepiness are presented here.

But first, it is important to distinguish two concepts, that, even though similar in how they are experienced, their physiological origin differs: drowsiness and fatigue.

While drowsiness results from the physiological need to sleep, mostly modulated by circadian cycles and hours slept in the previous nights, fatigue is associated with the tiredness resulting from performing a certain task. It is important to distinguish both, because different measure systems are needed to assess each correctly, and different strategies can be applied to them. As an example, time on task is a good simple measure to estimate fatigue, but fails to describe the individual's need to sleep. And while performing mentally engaging tasks, as simple algebraic calculations with speed sign values, seems to reduce the fatigue progression, it has no effect on drowsiness, which can only be resolved with actual rest [38].

2.5.1 Epworth Sleepiness Scale

This is a drowsiness scale designed to detect abnormal daytime drowsiness, which can then alert for the existence of an sleep disorder. It is based on a questionnaire asking patients to classify the likelihood of them falling asleep in 8 different situations in a 4 point scale. Those items are based on scenarios that occur occasionally during a normal day, but not necessarily every day. It is important to note that Epworth Sleepiness Scale (ESS) does not measure one's level of alertness/drowsiness at a given moment, but the average sleep propensity across different activities. For this reason it is used to screen sleep disorders, as it identifies individuals that live their day with excessive daytime sleepiness, but its unable to determine which causes or factors produce that state. [39]

2.5.2 Johns Drowsyness Scale

The Johns Drowsiness Scale (JDS) uses analysis of eye closure to determine in real time the level of drowsiness of an individual. It was created and calibrated for a specific eye tracking device, the Optalert[2], and has the advantage of being an objective measure of drowsiness, this is, it does not depend on the subjects capability to self assess their state. [40] By using the ratio between amplitude and velocity of blinking movements, this measurement does not need individual calibration. The major limitation of this scale is that it needs an infrared camera tracking the eyes to capture the dynamics of blinks. The scale measures drowsiness with a value from 0 (=very alert) to 10 (=very drowsy).

2.5.3 Karolinska Sleepiness Scale

Karolinska Sleepiness Scale (KSS) is a subjective scale of drowsiness, where individuals are asked to rate their state from 1=very alert to 9=very sleepy [41]. It consists on a single item self report measure, where the individual states which value (between 1 and 9) better correlates with its perceived state, after being explained to him that the levels span like the following:

- 1: Very alert
- 3: alert
- 5: neither alert nor sleepy
- 7: sleepy (not fighting sleep)
- 9: very sleepy (fighting sleep)

The ease of application of this scale made it popular in several drowsiness related studies, namely naturalistic driving [42–46], shift work [47–49] and attention and performance [50–53]. A study from 2006

[2] https://www.optalert.com/why-optalert/science/#johnsdrowsinessscore

by K. Kaida *et al.* [54] focused on the validation of KSS against EEG signals. The researchers add their contribution to other works confirming the validity of KSS to measure subjective sleepiness [42, 50, 55]. To do so, 16 participants were tested for 3 days, with repeated sessions in low lighting levels, where the KSS score was measured along side various EEG features, as alpha and theta wave power during a Karolinska Drowsiness Test, and the alpha attenuation coefficients. Behavioural components were also obtained, as the number of lapses and reaction times during a Psychomotor vigilance test. The results further confirmed the positive correlation KSS scores have with subjective sleepiness, finding almost linear relations between this scale and the physiological parameters gathered. To facilitate drowsiness detection based in this scale, it is common to aggregate ranges of KSS values into categorical labels that better differentiate different states of sleepiness, as shown in figure2.5 [56, 57].

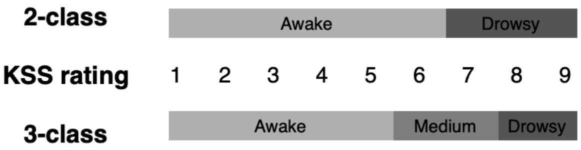


Figure 2.5: KSS ratings and corresponding proposed categorical labels. Adopted from Oliveira, 2018.

2.5.4 Stanford Sleepiness Scale

This scale is symmilar to KSS, in the way that it is a subjective measure of sleepiness. Individuals are asked to rate their sleepiness as one of seven states [58, 59]:

- 1: Feeling active, vital, alert, or wide awake
- 2: Functioning at high levels, but not at peak; able to concentrate
- 3: Awake, but relaxed; responsive but not fully alert
- 4: Somewhat foggy, let down
- 5: Foggy; losing interest in remaining awake; slowed down
- 6: Sleepy, woozy, fighting sleep; prefer to lie down
- 7: No longer fighting sleep, sleep onset soon; having dream-like thoughts

It is to note that this scale has its levels more detailed than the Karolinska one, which is less defined but allows intermediate classifications. The other main difference between these two scales is that Stanford Sleepiness Scale (SSS) has items that can be answered based on boredom or task related fatigue, while KSS keeps its focus on the propensity to fall asleep, thus being a better measure of sleepiness alone [60].

2.6 Decision Models

To model the relationship between HRV features and drowsiness, machine learning algorithms will be implemented, so serves this section as a brief review of the basic concepts that this field attains, as well as some of the techniques used to improve their results.

As part of the larger field of Artificial intelligence, machine learning comes from the need to have computers refine their functions, learning from provided data how to better solve given tasks [61]. As such, different strategies can be used to find this learning capability, in the majority of cases, a function is defined so that a set of its parameters can be adjusted by the computer in order to better emulate the real relationship between the inputs and outputs present in the data. This function can be as simple as a linear one, with only two parameters (slope and bias), or as complex as the highly non-linear result of a neural network with several hundreds of parameters to tune. And the model's dependency on data dictates that its that source material that defines how complex the model should be: complex models on simple data end up paying too much attention to noise, jeopardizing the learning procedure, and too simplistic models fail to fully reconstruct more complex relations in the data [62, 63].

To help find the equilibrium point between robustness against noise and flexibility to model complex data, different procedures can be implemented, such as tuning hyper-parameters (model parameters that are not defined during the training progress, but that are defined *a priori* and can have great impact on its capability to generalize) and selecting subsets of features to evaluate, besides the very first step, choosing adequate model architectures.

In the task this project is interested in, classification of alertness/drowsiness state from HRV features, the outputs of the data, and models, correspond to a finite set of labels or classes, more specifically two: 0=alert and 1=drowsy. Because it is a classification task, the models will define a separation boundary in feature space that places each provided data point in the correct class, measuring its performance and refining it in accordance to a cost function. The use of a cost function leads the model to tune its parameters in order to minimize its value, which ultimately corresponds to correctly classify the most samples possible. While in tasks where each class is completely separated from the other this results in perfect classification of all samples, in those cases where the classes distributions overlap, the model will have to place its border here the majority of samples is correctly classified. This means that if the classes are not balanced, the model can choose to classify every sample as the majority class, as that gives it the lowest error. To avoid this, class balancing techniques must be employed [64].

2.6.1 Models

2.6.1.A SVM

Support Vector machines are classifiers that search for separation hyper-planes in feature space that maximize the margin between classes. This strategy allows higher levels of confidence when classifying new samples that lie close to the decision boundary, as it was chosen in order to balance the space separation. Two major parameters to tune in this model are the kernel and the regularization parameter. Kernel, i.e. the function that defines the hyper plane geometry, and that can be used to model feature spaces that are not linearly separable, but perhaps are radially separable, using radius-based functions as the kernel. And the regularization parameter weights how much misclassifications cost to the loss function. This is especially important in cases where classes are not completely separable, this parameter allows a trade-off between margins and misclassifications, having low regularization parameters creating separation landscapes with higher margins and some misclassified samples, while larger values of this parameter favor the correctness of classification, sacrificing margin area. [65]

This type of model is largely used in classification tasks involving HRV due to its ability to store compact representations of feature spaces, even in non-linear cases. [66–69]

2.6.1.B One Class SVM

The model itself is the same as the SVM explained previously, but here, instead of modeling a separation boundary between two classes, the model is trained only with samples from one class, learning a self-contained area where that class lies on, and classifies as outliers all samples that fall outside of that region. Deploying this strategy on drowsiness classification is based on the assumption that alert samples constitute a normal class, and that the HRV get anomalous in drowsy states, producing this way outlier samples that can be classified as such.

2.6.1.C Gradient Boosting Trees

Gradient Boosting Trees are a classification algorithm that instead of building a single powerful representation of feature space, relies on multiple weak classification stages, each contributing with a different insight on how classes can be separated. By organizing these stages on decision trees, optimized contributions can be added up to form classifiers that are robust, even in the presence of noisy data [70].

2.6.1.D Artificial Neural Networks

ANN are networks of simple computing cells that try to mimic the process by which information is processed in the brain, a cell collects a series of inputs, and compares its weighted sum with a threshold to decide on whether to activate or not. Making use of alike computing entities, and tunning the weights by which inputs are multiplied, artificial networks can approximate any function. Though this universal approximator is extremely useful in non-linear classification tasks, the number of parameters that the model needs to adjust to perform well in such tasks demands an increased amount of data to successfully train a neural network [71].

2.6.2 Feature selection

Feature selection is the process by which the total set of available features is reviewed and possibly trimmed. Today, when collecting information, given the apparent limitless of storage space, the tendency is to accumulate every possible feature available, however, when that same data needs to be processed and utilized by a machine learning algorithm, the high dimensional feature space is more often than not a curse. As more features does not necessarily reflect more information for the models but implicate heavier memory usage for feature space representation and longer training times, methods to select only those features that contribute with relevant and unique information were created [72]. There are different strategies to select an optimal set of features, with two sitting on opposite extremes: supervised and unsupervised. While supervised directly connects feature selection with model optimization, and evaluates each possible subset of features based on the model performance it results in, unsupervised feature selection tries to evaluate the vectors in feature space independently from classification labels or outcomes. Two major quantities are considered in this process: relevance and redundancy. The first deals with a measure of how rich can be the information encapsulated in each feature, and deals normally with dispersion measures, as Variance or Mean Absolute Deviation (MAD). Redundancy represents a quantification of how independent the information given by two features is, being that high levels of redundancy encourage the elimination of one of them. This is measured using metrics of similarity, as correlation coefficient or angle cosine between feature vectors [73]. By ranking features based on their relevance, and iteratively adding features that present low redundancy with the already selected ones, a subset of features that maximizes information maintenance, while keeping only those that contribute with unique knowledge, is created [73].

2.6.3 Class balancing

From the various factors that hinder machine learning algorithms performance, a commonly reported one is the imbalance between cardinally of samples in each class. While, as pointed by Batista [74], class balancing itself does not pose a problem, when other issues co-exist, as class overlapping, performance is lowered. Imbalanced classes occur frequently in real classification tasks, specially when trying to detect uncommon but important events. If the feature space allows complete separation between majority and minority classes there is no effect of this imbalance on performance, but, if classes overlap and the model is forced to choose a separation plane that misclassifies a few samples, the minority class tends to be sacrificed in order to minimize the cost function, most of the times compromising the detection of that important infrequent event. Batista shows in his work that oversampling improves model performance in these cases. One of the methods used to over-sample the minority class is Synthetic Minority Oversampling TEchnique (SMOTE), which has the advantage that it generates new unseen samples, unlike random over-sampling that simply duplicates the original ones, potentially leading to over-fitting of the model. SMOTE produces new unseen samples by selecting the k-nearest neighbors and adding to the dataset points that lie between the original point and a fraction of its neighbours, depending on the oversampling needed. This balances the cost of misclassifying minority samples, and generalizes the feature space region occupied by this class, improving its detection accuracy [75].

2.6.4 Feature space visualization with t-SNE

The first difficulty encountered when dealing with data with dimensionality higher than 3, is that it becomes impossible to produce a direct visual representation of feature space. This visualization allows quick inference of data characteristics, as linear relationships or existence of clusters. One tool that is particularly praised for its capability to reduce feature space dimensionality and reproducing the clusters in the new visible space is t-Distributed Stochastic Neighbor Embedding (t-SNE). By converting euclidean distances into conditional probabilities that represent similarity, this method is capable to map the points into a lower dimension space in such a way that it maintains the local structures present in the high dimension data. For this reason t-SNE is the tool of choice to have a visual representation of high dimensionality feature space structures [76].

3

Systems for stress and fatigue detection - state of the art

Contents

Automobile field is one where the need for insight on a persons internal state is becoming already a main topic among researchers. Through the need to reduce road accidents and the unavoidable evidence that most incidents are caused by drivers. According to a report from the Department for Transport [77] (UK) in 2008 3% of fatal accidents and 2% of those that result in serious injury had fatigue as a contributory factor. However, previous research pointed out larger number, namely 10% of collisions [78] and 17% of road crashes [79] that result in injuries and death being sleep related. Its is plausible that fatigue as an accident cause is underestimated in official reports, given the lack of specific formation of police agents to assess its contribution when reporting accidents and the fact that drowsy drivers involved in crashes tend to be wide awake when interviewed because of the induced stress [80]. This underestimation predicts that a much more realistic statistic would be that around 20% of all road accidents are caused by fatigue, either by actually falling asleep on the wheel and or by the decreased performance that it implies.

According to an European Road Safety Observatory report from 2018 [81], driving while sleepy or fatigued has a prevalence much higher than what would be expected or even minimally safe, surveys demonstrate that more than a half of the population drives while being drowsy at least once a year, with a range of 10%-40% of them having actually fallen asleep on the wheel. Also, studies from the united states corroborate these results, as about one third of the population feels impaired to perform their daily tasks at least once on a monthly basis [82], which included severe reduction in driving performance. The same report states that fatigue related accidents result in high level injuries, and reaffirms the 20% prevalence of fatigue as a crash contributor. Finally, different studies focused on measuring the increased risk resulting from driving while drowsy, finding the risk to be involved in a car crash to be 4 to 14 times higher than for rested individuals [83–85].

Another aspect of driving behaviour that is interesting to analyse is road rage. The American Safety Council defined road rage as ”an attack initiated by the driver of a car or a passenger, on a driver of another car or its passenger, using a car or another dangerous vehicle, this anger being the result of an incident or event on the road during driving” [86]. And such attacks can take the form of tailgating, blocking passage for other drivers or even verbally and physically assaulting them. As these are risky behaviours to present on road, a study dedicated to establish a relation between perceived stress on road and driving aggressivity [87]. By observing the behaviour of 226 drivers entering a parking lot with heavy traffic, annotations on aggressive behaviour were taken, and a set of questionnaires was used to assess the drivers coping (with stress) style, general driving behaviour and stress state during the experiment. From these measurements it was observed that in general, the more stress is perceived by a driver, the more aggressively he behaves. However, other factors model this, as intrinsic coping mechanisms, for example, drivers with problem-solving coping strategies tend to present less aggressivity when driving than emotional ones. These results are similar to those obtained in previous studies [88, 89], that directly related stress with poorer driving performance, with the observation that, if on one side, aggressive-coping stressed drivers tend to overtake other vehicles more often and in a more error prone manner, while non-aggressive, but driving disliking drivers tend to be more cautious even though they presented less control. Moreover, following a similar procedure, but coupling it with measurement of reaction times, Różanowski [90] established a positive correlation between perceived stress and poorer task performance in driving environments, which was even more evident in aggressive-coping drivers.

Seeing this, it becomes clear that a need for measures that reduce these human factors preponderance in road accidents is increasingly important. The most obvious course of action would be to remove the human from the driving process, which is already a market direction in the form of autonomous vehicles. However, full implementation of this technology will not occur in the next 30 years, which means that other strategies are needed [91]. Those are the creation of systems that monitor and act upon the stress/fatigue state of the driver. This is advantageous because the only question to be solved is how to measure the internal state of the individual. Mainly three different branches on this subject appear:

- **Driving behaviour -** Analysing driving patterns, as angle of steering wheel corrections, lane positioning, etc.
- **Tracking face position and eye direction -** The position and movements of the head and eyes present characteristic patterns depending on fatigue/stress state.
- **Physiological signals -** Physiological manifestations are the direct result of internal state variations in an individual.

Several proposals have been produced in this scope, some of which are presented in the following sections.

3.1 Driving behaviour

In 2002, Kirche *et al.* published an article that evaluated a series of existing technologies for fatigue detection through evaluation of driving patterns [92]. The authors describe methods directly related with the steering wheel movements and steering wheel angle variability. According to Wylie *et al.* [93], steering wheel variability can be related with driver drowsiness because there are minimal frequent and low amplitude route corrections during a driving session, that tend to reduce in number and increase in amplitude as drowsiness evolves. Although easy to extract this type of information, geometry of the road affects this greatly, as the distinction between small heading corrections and simple following of road geometry is affected not only by this geometry itself, but also by the velocity of the car can change the temporal span of these phenomena. Eliminating the average of the steering wheel position over a defined period of time/ length of road is often used to reduce the weight of road geometry in this set of parameters, however, in urban environments the length, duration of curves are highly confoundable with micro-corrections, thus these methods are relevant only for highway studies. One other method is referred in [92], that is the computation of a VHAL index, explained in detail by Bittner [94]. The *VHAL index* corresponds to the squared derivative of the position of the steering wheel, in that paper denoted as *HAL*, so that it can be thought as a measure of its variance. By band-passing the calculated squared derivative of HAL, a relatively smooth measure of the variability of steering wheel movements is obtained, which seems to be related with drowsiness as when it increases, the drowsiness level decreases and vice-versa.

One other set of commonly used parameters is related with the lateral position of the car in relation to the lane limits, the capacity to maintain a steady, well centered position is a sign of alertness.Dingus *et al.* [95] compared lane positioning-related features with PERCLOS (percentage of time with eyes at 80% closure) to evaluate their suitability to detect drowsiness. One other work that searched for more of these features was conducted by Skipper [96] in 1984. Together, these two articles propose lane position features as standard deviation, mean square and maximum of lane deviations, as well as measures of lane deviation that are heavily weighted for lane exceeding or mean square of high passed lane deviations, where all of them showed to be highly correlated with eyelid closure, and thus drowsiness, except for the last one, where some potential was identified but not validated as thoroughly as the others.

One final set of parameters covered in Kircher's work [92], is the time-to-line crossing, which corresponds to the time it would take for the car to exceed the lane limits if its instantaneous lateral velocity is kept. This is a rather difficult measure to obtain, as no direct sources of this velocity information are available, so approximation methods have to be deployed. Normally obtained from a first and second derivative of lateral position, measures of time-to-lane crossing has been correlated with drowsiness by various researchers [97–99].

In 2017, spanish researchers Muñoz-Organero and Corcoba-Magaña [100] combined measurements of car velocity, acceleration and Positive Kinematic Energy (PKE) with LF/HF ratio from HRV to predict upcoming values of stress, being actual stress measured as the aforementioned ratio. This approach showed some promise as a model trained and tested by the same driver would achieve accuracies around 97.5%, showing that kinetic parameters are also relevant for this type of driver evaluation.

3.2 Face/eye tracking

A lot of information can be extracted from the face and eyes of a person. As humans, we are hardwired to perceive immediately the internal state of an individual just by looking at his eyes. However, while it is naturally easy for us to detect ”sleepy eyes” or understand head tilts and different body poses as drowsiness, their definition for automatic detection of this state is not trivial to compute. Because of this, much work has been done in order to develop reliable and efficient methods that track the desired features, as well as to find those features that better describe the individual’s internal state.

Ji and Yang published in 2002 an article [101] proposing an image processing system that tracked in real time the pupil to determine the Percentage of Eyelid Closure (PERCLOS), Average Eye Closure Speed (AECS) and direction of gaze, all of which provided very encouraging results for drowsiness and alertness detection, identifying clear patterns that distinguished each situation.

PERCLOS is defined as the percentage of time that the individual has his eyes closed during a certain period of time. it is hypothesised that as a person gets drowsy, its ability to keep the eyes wide open deteriorates, so that higher levels of PERCLOS are observed in this situation [102, 103]. As an example, Ji and Yang noticed that PERCLOS above 30% were associated with drowsy states [101].

AECS is the average time it takes for the eye to go from fully open to fully closed, and the other way around, measured over a regular time window. Research shows that drowsiness results in slower eyelid movements, such that noticeable increase in AECS happens [101].

In terms of direction of gaze, Ji and Yang found that lack of attention was related with lateral deviations (to the right or left), while drowsiness resulted in its lowering [101].

In 2006 Bergasa [103] produced a similar study, where some parameters were added, namely the frequency of nodding and blinks, as these are clear signals of tiredness. However, the difficulty involved in tracking the head in 2D pictures and resolving a three dimensional motion for nodding and the camera frequency needed to detect fast blinks render these measures less feasible than those already described. One other parameter tested was a detection of fixed gaze, which means that the direction of gaze remains constant for a prolonged period of time. When the driver is drowsy or inattentive, his gaze remains still because there are no new *stimuli* to change his focus of attention, contrarily, active drivers continuously move their gaze around a restrict are to evaluate different cues present in the road. In the same year, Smith *et al.* [104] published an article describing another system, very close to these, where eyes, lip corners and face contours were detected to track eye closures and direction of gaze, and use both metrics to warn the user of low visual attention to the driving task. Masala and Grosso in 2014 [105] proposed a system robust to light changes and different users by evaluating images through a dictionary of poses and a dissimilarity classifier, where the dictionary of poses is based on the same fatigue-pose relations that were described already. And Zhang *et al.* [106] used the same fatigue models, but RGB-D cameras to ease the three dimensional reconstruction of pose problem.

3.3 Physiological signals

While the previous measures relate to the manifestations that are observable when drowsiness sets in, and that are easily interpreted by humans, they lack to detect the deterioration of alertness of the driver at its origin. Measuring the physiological signals that are related to the very same processes that produce such cognitive state may be harder to accomplish and extract meaning from, but, if done properly, provides a window to detect subtle clues that otherwise would just be visible in later drowsiness states, and possibly, too late.

One of the signals that is agreed to present the best correlation with fatigue and drowsiness states [107], as well as with attention vs. inattention is the Electroencephalogram (EEG) [108]. The EEG measures the electrical field produced by neurons in the cortex, being possible to identify frequencies of neuron activation with temporal and (coarse) spatial resolution, frequencies that provide the needed information for the aforementioned correlations. Because of this, different researchers have proposed systems that monitored EEG to detect drowsiness or inattention of the drivers [109–112]. However, these systems impose the wearing of a set of EEG electrodes, that are a rather intrusive form factor, resulting in low adherence by users, and thus its impracticability in real life use.

Other signals can be used to overcome this, like Galvanic Skin Response (GSR), that is a good measure of the state of arousal of the individual, and ECG and PPG to extract the cardiac rhythm.

GSR measures the small changes in skin electrical conductance resultant from the activity of sweat glands, which are modulated by central autonomous activity [113]. While capable of accurately detect stress and drowsiness episodes, as shown by Healey and Picard [114, 115], and described by Chowdhury *et al.* [116], it is also very sensitive to ambient temperature, as high temperatures dominate the sweat glands behaviour over parasympathetic activation.

At last, two signals that convey the same fundamental information, and that seem to circumvent the limitations of EEG and GSR are the ECG and PPG, from which the dynamics of cardiac rhythm can be assessed. These signals are probably the easiest to integrate in an Advanced Driver Assistance System (ADAS) implementation, as ECG can be collected through any system with two contacts, like a chest strap, a tight t-shirt with contacts, or even a steering wheel with conductive leather. PPG is even simpler to integrate, as the wrist band and smart watch form factor can seamlessly enter the daily lifestyle of practically any driver. Driver state information is inferred by analysing the heart rate variability, and several methods using it have been proposed [57, 100, 114, 115, 117–120]. From these studies, superiority of ECG signal quality has been reported, justified by the normal usage of chest straps, that guarantee a robust fixation of contacts with skin during the all driving time, while wrist worn PPG sensors have shown sensibility to movement. Healey and Picard [114] achieved accuracies in the detection of stress with HRV features around 52.6%, referring the difficulty in establishing a defined boundary between stress and no stress as one of the main causes of this low performance, while Muños [100] managed to achieve 97.7% accuracies in predicting stress levels by using a regression to the LF/HF ratio. Gruden [120] encountered statistically significant differences in HRV features for different levels of cognitive load . And Silveira and Oliveira [56, 57], in different works, found that HRV features could detect alert states with 84% accuracy, but drowsy ones only with around 47%. This lower performance in drowsy states can be related with the continuous process of accumulating sleepiness, leading to overlapping classes and a blurry boundary between them.

In several the studies presented in this chapter, more than one information source was studied simultaneously, with those studies agreeing that ultimately, the combination of different signals was the best strategy to cover multiple dimensions of drowsiness and stress, and so capture the real state of the driver.

Table 3.1: Strategies found in literature to identify driver state through HRV features.

Source	Year	Model	Features	Results
[120]	2019	-	HRV - time domain	Statistically significant difference between cognitive states
[57]	2019	SVM, GBT, RF	HRV - time and frequency domain	84% accuracy for alertness, but only 47% for drowsiness
[56]	2018	SVM, GBT, RF, KNN	HRV - time and frequency domain	84% accuracy for alertness, but only 47% for drowsiness
[100]	2017	Regression	HRV - LF/HF ratio	97% accuracy for stress
[114]	2000	Linear Discriminant	HRV - LF/HF ratio	52.6% for stress

4

Methods for signal extraction

Contents

To later implement a drowsiness detection system, three different signal sources are tested, each with its particular interface with the user, and its output signal characteristics. To ease the integration of all these devices into this workflow, specific strategies for output normalization were implemented. Depending on the signals needs, filtering and further processing were applied to ensure that all devices return a trustworthy IBI sequence.

4.1 CardioWheel

CardioWheel was originally developed as an off-person biometric solution based on ECG. Lourenço *et al.* built a PCB designed to process one lead ECG signal, filtering it with a band pass filter and segmenting the successive heart-beats. While the segmented signal is used to create templates of mean ECG waveform for individual identification, the system also produces a stream of time intervals between identified R peaks [121]. The system also implements an outlier detection mechanism, that discards noisy or abnormal segments [122], and a hands-on-wheel detection system that switches off the R peak location when the user is not properly contacting the sensors.

By having conductive leather in a steering wheel cover, and integrating that sensor form factor with the Printed Circuit Board (PCB), a non intrusive method for ECG collection is obtained, which can be easily introduced into the driving environment.

As CardioWheel already provides a stream of IBIs, that is the signal used in this work, only passing it through a IBI revision system before performing HRV analysis.

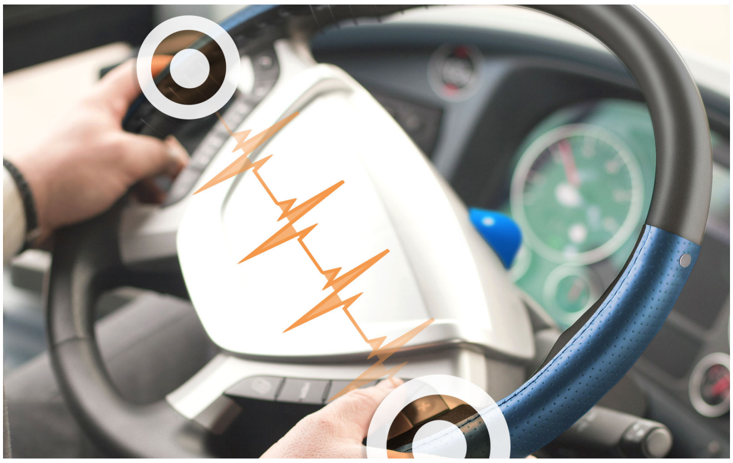


Figure 4.1: CardioWheel

4.2 Wrist PPG

Advances in wearable technology, such as microprocessor computing power, battery life-time and device miniaturization, as well as signal processing techniques designed for PPG, present us today with a long desired possibility: continuous recording of physiological signals like the cardiac rhythm with minimal intrusion on peoples lives. Current form factor used for these solutions, the smart bands and smart watches are even considered by the general public as trendy, so it can only be expected that the penetration of this type of technology will continue to increase during the following years [123].

However, current usage of this type of technology is limited to rough estimations of heart rate averaged over a time period (as one minute), to satisfy the curiosity of some and, in more applied users, measure and control sports performance.

This limitation comes from the fact that, being a non-intrusive optic based method, PPG greatly suffers from motion artifacts and others resultant from sudden changes in ambient light. To eliminate those, several types of filtering and machine learning methods have been presented, but, given their computational burden, they are mostly unsuitable for these mobile solutions. Because of this, most cases couple an accelerometer that, when movement is detected, HR determination is suppressed, as bad signal quality is assumed in these situations.

In this work, the device used for processing and algorithm design was the PulseOn hBand (fig. 4.2), a wearable wrist band with two PPG sensors using green LEDs, collecting raw data at 25Hz. Other devices were tested, namelly the Maxim MAXREFDES103, and the Emotibit. All of them provided the same sampling rate, and while Maxim's device had built in algorithms for HR and IBI detection, and Emotibit offered a more flexible and programmable framework, PulseOn wrist band was chosen. The use of such a device with low sampling frequency is motivated by the ease with which it is possible to collect raw data with this comfortable form factor, but also its battery autonomy and the fact that the sampling frequency is in accordance with most wearable wrist bands with PPG sensors.

4.2.1 Signal treatment

Raw PPG data is a composition of real PPG oscillations, noise and artifacts provoked by hand gestures and ambient light changes that affect either the contact the sensor has with wrist circulation or the overall luminosity in the sensor-wrist environment. For this reason it is necessary to reduce the weight that this undesired components have (figure 4.3).

The first step was to eliminate all light fluctuations that are slower than human heart rate and light offset changes, so a moving average with window corresponding to 1 second of signal is removed for three iterations. To guarantee the possibility of implementing this system on a online assembly, the necessary coefficients for digital filtering were determined through (4.1). This results in a zero-

Figure 4.2: PulseOn hBand Wrist band.

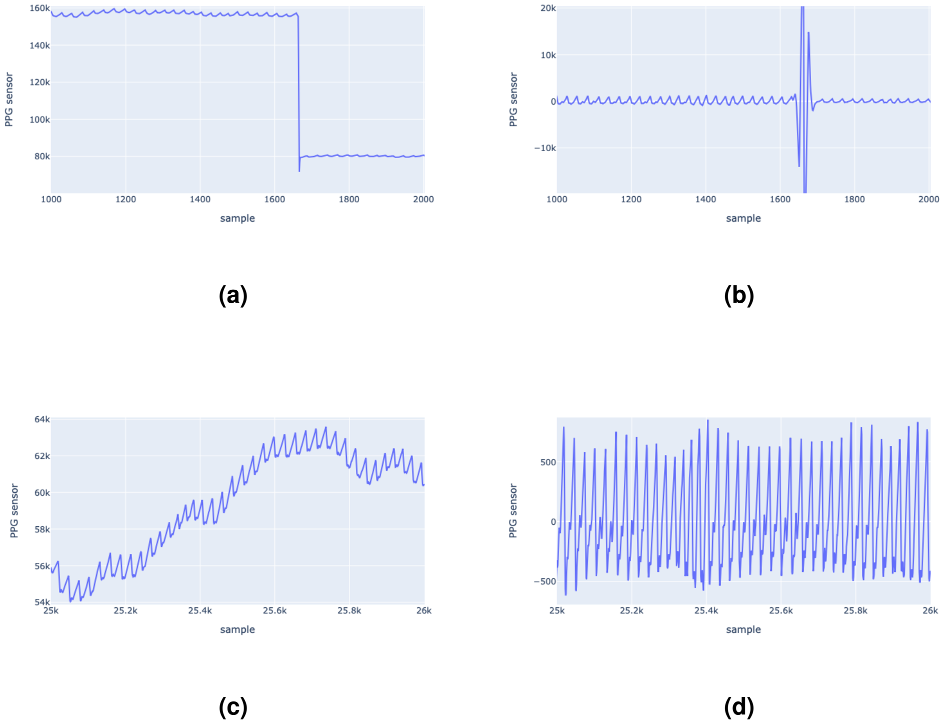


Figure 4.3: Filter results: (a) raw signal with discontinuity; (b) result from filtering (a); (c) raw signal with wandering mean; (d) result from filtering (c)

mean oscillatory signal where PPG pulses are evident enough to be easily distinguished from remaining artifacts.

$$\begin{aligned} y[n] =& x[n] - \frac{3}{W}\sum_{k} x[n+k] + \\ &+ \frac{3}{W^2}\sum_{k}\sum_{v} x[x+k+v] - \\ &- \frac{1}{W^3}\sum_{k}\sum_{v}\sum_{u} x[n+k+v+u] \end{aligned} \tag{4.1}$$

Where W is window, that should equal the sampling frequency of the signal, and k,v,u go from -W/2 to W/2, to center the window on the point being evaluated in each moment.

Most of the remaining artifacts resemble peaks, but have amplitudes and locations that can identify them as outliers from the set of normal peaks. A good peak detection algorithm should discard these false peaks.

4.2.2 Peak detection

Peak detection with adaptive threshold inspired on Thang's paper [18] was implemented to identify systolic peaks (figure 4.4).

Adaptive threshold consists on a method to keep track of successive peaks where a dynamic threshold alternates between two modes to identify relevant *maxima* of an oscillatory signal.

The first mode consists of following the waveform until a peak is reached. The first detected peak has to have an amplitude above 200 (defined empirically) and be the global *maxima* in a 8 sample (320 ms) window. After this, a slope is defined and the threshold value changes with time independently until it crosses again the signal, according to equation 4.2.

The slope is altered every time a new peak is found, to adjust it regarding peak amplitude and current cardiac rhythm. When a peak is missed, *i.e.* no peak is found 1.5*last_IBI after the last found peak, the peak detector is reset so it is able to detect peaks in a different regime (different amplitude or shape configurations). The threshold updating for successively found peaks was made using the expression depicted in equation 4.3. The so defined slope guarantees that even peaks in fast heart rates that decrease to 50% of the previously found amplitude are detected and accepted as valid, making this process robust to changes in wave amplitude and to different cardiac frequencies.

Also, to guarantee that sudden peaks provoked by motion artifacts or other noise superposition does not introduce false peak detection in impossible locations, a refractory period was established. Initialized as 500ms (12 samples) advance in sample acceptance, as soon as the first inter-beat interval (IBI) was

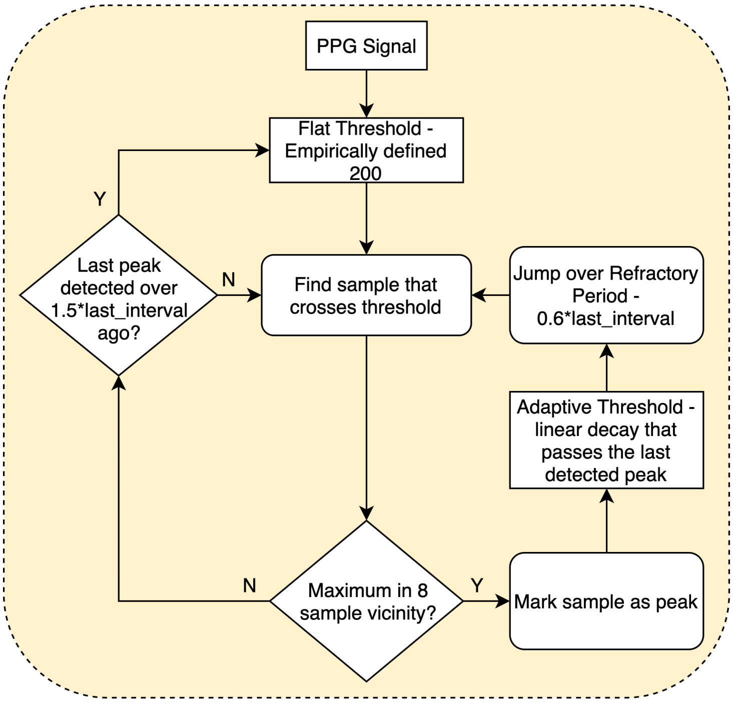

PPG Signal
Flat Threshold - Empirically defined 200
Y
Last peak detected over 1.5*last_interval ago?
N
Find sample that crosses threshold
Jump over Refractory Period - 0.6*last_interval
Adaptive Threshold - linear decay that passes the last detected peak
N
Maximum in 8 sample vicinity?
Y
Mark sample as peak

Figure 4.4: Flowchart of the adaptive threshold algorithm used for PPG peak detection

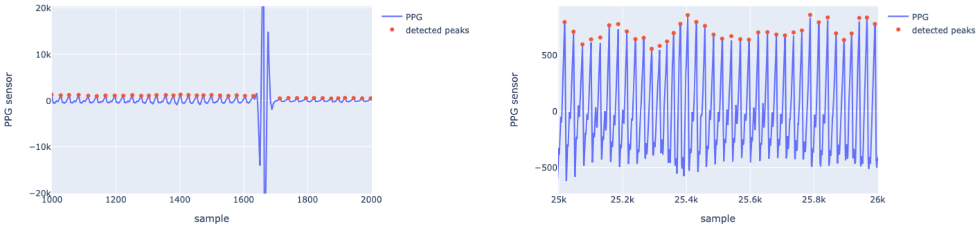
determined, the refractory period started to be updated according to 4.4.

$$Thrs[n] = Last_y - slope \cdot (n - Last_x) \tag{4.2}$$

$$slope_{n+1} = -\frac{0.5 * last_y}{F_s * IBI_n} \tag{4.3}$$

$$RP_{n+1} = 0.6 \cdot IBI_n \tag{4.4}$$

After this independent threshold crosses the signal again, the following mode is recovered until a new peak is found and a new threshold and RP are defined. The dependency of each slope on the previously found peak amplitude makes this method robust against amplitude variations in successive peaks.



(a) **(b)**

Figure 4.5: Peak detection results, here the same are shown the same segments as the ones represented in filtering section.

4.3 Movesense

Movesense is a lightweight sensor, that can measure ECG. Connected to a chest band, it is able to monitor a persons heart signal with great Signal to Noise Ratio (SNR), being the collected signal robust against motion or contact artifacts, due to the elasticity of the band. The use of this equipment serves as a source of groundtruth signal for heart dynamics, against the CardioWheel and the wrist bans, as it measures it directly in its source.

Figure 4.6: Movesense chest band.

As the signal obtained through this device has a very good SNR, figure 4.7 represents the signal quality observed throughout all measurements made, no filtering is implemented, and a simple Pan-Tompkins algorithm [124] was used to detect the R peaks, whose location was revised by selecting the closest *maxima* from the original estimates.

From the detected R peaks, IBI values were calculated, which, after being revised by the IBI corrector, were used for HRV analysis.

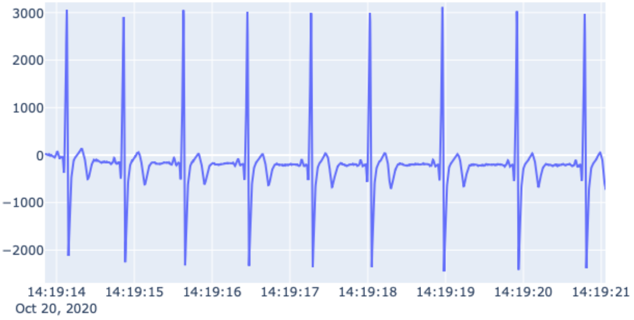


Figure 4.7: Example of ECG signal collected using the movesense chest band, the high quality of this signal allows R peak detection without firstly filtering it.

4.4 IBI corrector

It is a common practice to filter IBI values before performing HRV analysis, as outliers can deviate the variability indexes from their true value, hindering any further conclusions about the recorded signals. In most cases, IBI revision is made by simply eliminating non-physiological intervals, such as the ones shorter than 300ms (above 200 Beats Per Minute (BPM) [125]) and those longer than 1500ms (below 40 BPM [126]). Other approaches even define boundaries to how much consecutive IBIs can differ, discarding those that cross so defined thresholds. However, while this outlier elimination strategy improves results in conventional HRV time windows, where several hundreds and thousands of IBIs are available, windows as short as two minutes may not be able to afford the information loss by discarding outliers.

For this reason, this work proposes a system capable of not only identify outliers, but also of reconstructing the real IBI values from signal corrupted with outliers, combining the reliability of HRV measures based on only physiological values, but also maintaining all the available information, so that the analysis is not compromised by scarcity of data.

This system is based on the ratio between consecutive IBI. Outliers are defined as points which ratio crosses a defined threshold. This forms a basis to evaluate streams of IBI using the same sets of criteria, regardless of the absolute values present in any record.

To define the thresholds and test the performance of the corrector, ECG records from the naturalistic driving experiment SleepEye [127] was used. Signal from all available records was visually inspected, and all segments with clean and correctly identified R peaks were converted into series of IBIs. This way, a dataset of validated IBI streams was ready to evaluate the corrector, consisting of 138 intact series, containing 231470 consecutive pairs of IBIs in total.

Firstly, to define thresholds, a histogram of the ratios produced by the consecutive IBI values was produced, resulting in figure 4.8.

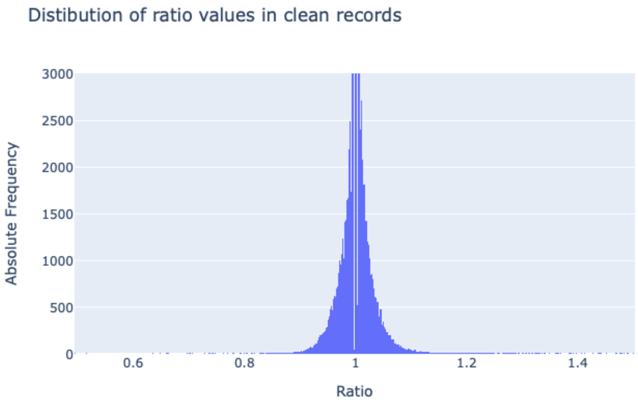


Figure 4.8: Distribution of IBI ratios

This distribution of values gives good grounds to select a lower threshold of 0.8, and a upper threshold of 1.5. These limits were designed to be in accordance with previous research indicating that variations larger than 20% between IBIs indicated outliers [128]. Ratios are also advantageous to allow identification of how many heart-beats were missed in cases of longer IBI. Rounding the ratio to the nearest integer would return the number of heart beats encapsulated in the same outlier IBI.

These defined thresholds identified only 0.02% (53) IBI as outliers.

4.4.1 Algorithm design

There are three main functions the corrector has to implement: to track the level of reliability of each new IBI, and, having identified an outlier, decide whether to fill a gap, our join two smaller intervals into a physiological value.

Four thresholds are defined, physiological bounds for IBI values, and ratio limits for normal inter IBI variation:

- I_{inf} - Inferior limit of physiological IBI, set to 300ms.
- I_{sup} - Superior limit of physiological IBI, set to 1500ms.
- r_{inf} - Lower bound of accepted ratio, set to 0.8.
- r_{sup} - Upper bound of accepted ratio, set to 1.5.

While physiological limits are used to make a final validation on accepted IBI, ratio limits are used in a function that transforms the exact quotient between IBIs into an integer (algorithm 4.1).

Algorithm 4.1: Definition of integer levels of ratio

```
Result: ratio
quotient=$IBI_n/IBI_{n-1}$;
if quotient< $r_{inf}$ then
    return 0;
else
    return round(quotient$-(r_{sup} - 1.5)$);
end
```

The rounding function used in algorithm 4.1 rounds floats to the nearest integer. This behaviour is used so that ratios larger than r_{sup}, there is a estimation of how many real IBI were skipped to produce the larger outlier. As an example, with $r_{sup} = 1.5$ and a quotient of 2.8, the system would be capable of realizing that most likely 3 IBIs, instead of only 2, were concatenated into a single value.

Figure 4.9 explains the logic the corrector implements. The main process consists of consuming a value of a waiting list of IBI, and, by deeming its corresponding ratio to the previously accepted value, decide whether to directly add it to the validated results, to fill detected gaps or to sum it to an adjacent interval.

Before starting this process, and any time the corrector needs to be reset, the corrector must initialize the `pending` list and the `last` value. To do so, the corrector extracts the first IBI from the input and checks if it is inside the physiological range. If so, that value is defined as `last`, and the rest of the available IBIs form the `pending` list.

To fill detected gaps, a series of estimates for the missing IBIs are computed, using the mean value between the partition of the longer interval and the last accepted value (eq. 4.5).

$$new = \left(\frac{Value}{ratio} + last\right)/2 \tag{4.5}$$

This is done to simulate a smooth evolution from the last accepted IBI and the partition length needed to have a detected heart beat at the timestamp corresponding to the longer outlier. It allows smooth shortening or widening of estimated intervals to accommodate outliers that are not integer multiples of the last valid value, instead of having a sudden jump to a series of identical partitions of that outlier, as seen in figure 4.10.

After defining a filling value, the outlier gets this estimated value subtracted from it, and is replaced at the beginning of the `pending` list to proceed the evaluation, if the remaining value continues to be large enough to be an outlier, the filling process is repeated.

Finally, if the system detects a shorter interval, it tries to join it with an adjacent value. This serves to correct instances where a false peak was detected, leading to a real IBI divided into two parts. The corrector chooses the smallest value between the previous (`last`) and following intervals, and adds the outlier to it. Finally, it checks that such addition does not result in a ratio above r_{sup}, if it does, the shorter

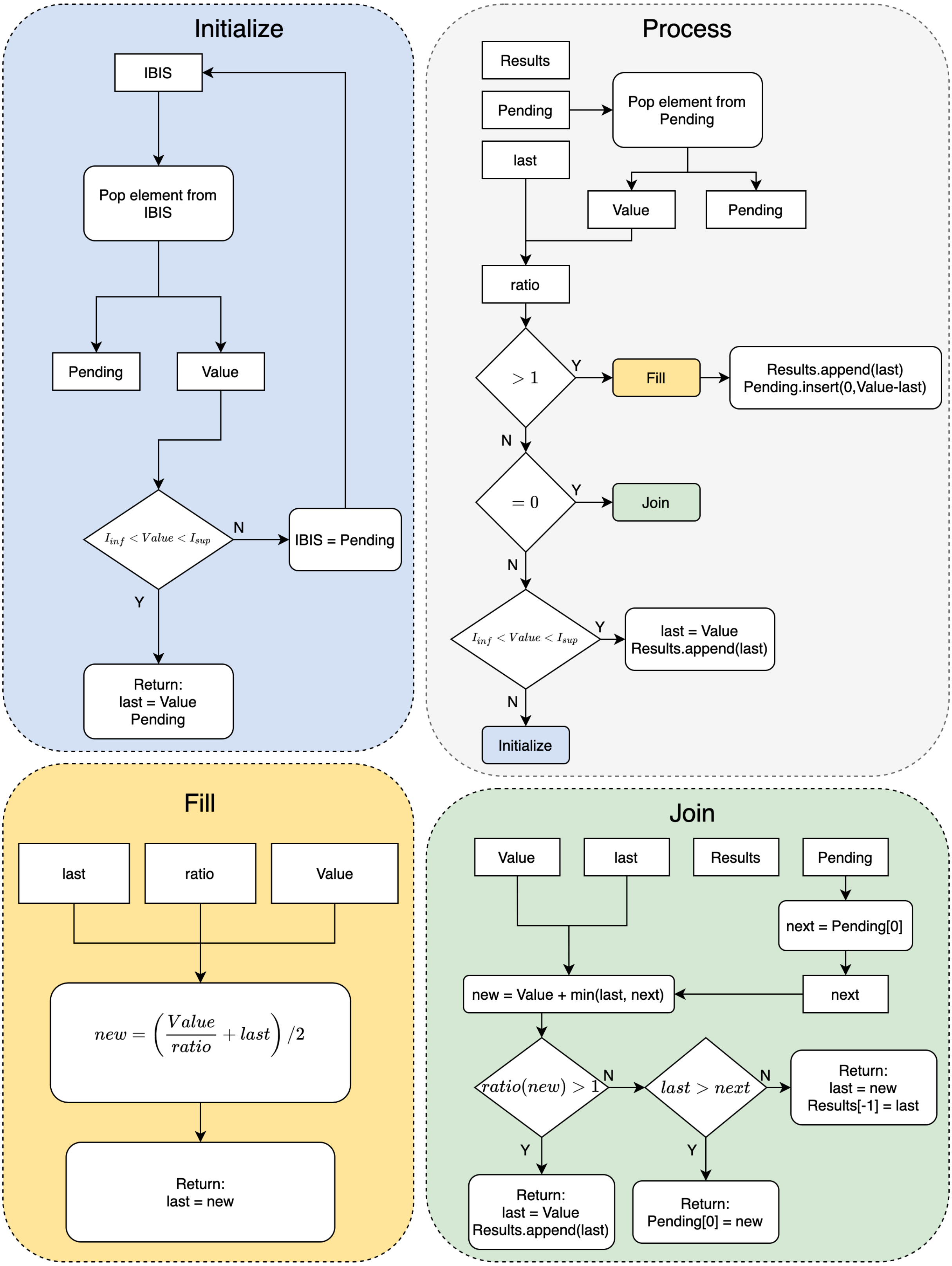


Figure 4.9: Flowchart of IBI corrector functioning, evidencing the main process and specific tasks: initialize, fill and join

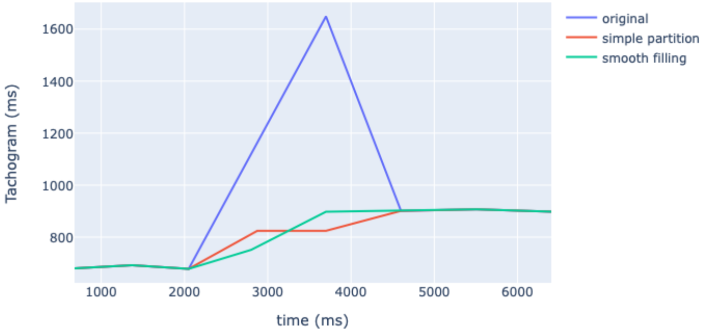


Figure 4.10: Comparison of missing IBI estimation using simple homogeneous partition of outlier, or the smooth filling strategy.

interval is added without any processing, as it would mean that it did not correspond to a partitioned IBI.

To make sure that the system always produces realistic estimates, in regards to physiology, any time a proposed value reports a normal ratio, but a non-physiological value, the corrector is reset by running the initialize process on the current `pending` list.

4.4.2 Performance Evaluation

To evaluate the performance of this system, the 138 clean ECG records described before were used. Artifacts were added to those signals, with controlled percentages off contamination by miss detections and false peaks. To form such signals, a desired percentage of IBI values were chosen at random, some to be joint with the following value, simulating a missed detection, and, thus, returning a longer interval. Other would be divided in two at a random proportion. Percentages of missed detection and false peak artifacts were controlled independently, and the generated error only made sure that the same original IBI would not be selected for both contamination types. To measure performance, re-sampling of the original and corrected tachograms was performed, from their irregular sampling to 16Hz, to measure the quadratic error of the correction. By contaminating all records with a defined level of artifacts, summing all the resulting absolut errors and storing the number or re-sampled samples allowed to calculate a value of MAD for the corrector at each contamination setup.

A grid of artifact density was produce, raging miss detections from 0.0 to 0.3 and false peaks from 0.0 to 0.1 in intervals of 0.05 and 0.02 respectively.

The results of such evaluation are presented in table 4.1.

To achieve these results, the threshold r_{sup} was changed from 1.5 to 1.7, as it enhanced the performance

Table 4.1: MAD of IBI reconstruction with different levels of contamination.

		False peak density					
		0.00	0.02	0.04	0.06	0.08	0.10
Missed detection density	0.00	0.13	0.20	0.30	0.58	1.56	1.99
	0.05	2.66	2.14	2.43	2.79	3.75	3.98
	0.10	5.23	5.63	5.97	6.62	6.20	7.42
	0.15	10.40	10.08	11.56	11.84	12.70	12.39
	0.20	15.96	17.27	17.64	17.19	18.35	19.17
	0.25	24.39	25.19	24.00	27.20	26.62	25.95
	0.30	35.93	36.19	36.84	35.79	37.37	38.03

of the corrector in clean ECG records.

These results sustain the decision to use this method, as the reconstruction of signal is almost perfect for signals contaminated with up to 5% of missed detections and 10% of false peaks. Errors in higher contamination levels are also acceptable, having in mind that the highest measured level corresponds to having only 60% of the original values available, and it is still lower than the uncertainty in IBI values produced by the PPG sampling frequency of 25Hz.

5

Detecting drowsiness from HRV features

Contents

To establish a set of models capable of correctly identify dangerous states of drowsiness, machine learning strategies were implemented, using a dataset from a previous naturalist driving study, Sleep-EYE [127]. This study consisted of 20 individuals who had their ECG measured and KSS self-report annotated during 90 minute drives in public roads in Sweden. Using these measurements, ECG can be transformed in HRV features and KSS annotations can justify a binary classification of alert/drowsy. Each individual drove twice, first in a day period, after a normally slept night, without influence of alcohol or caffeine, and the second in the night, after spending the day awake in normal activity. This measurement design intended to force alert and drowsy data from all individuals, even though it is not guaranteed that each individual record does not have a wide range of KSS scores associated with both alert and drowsy states.

The initial approach to the training of models using this data is inspired in a previous work that also tried to build classifiers for the task of detecting drowsiness from physiological data [117]. Using the same data set, the author tested four models: Support Vector Machine (SVM), Gradient Boosting Tree (GBT), Random Forest (RF) and Artificial Neural Network (ANN), comparing the performance between models and evaluating the effect that some alternative training strategies could have in their results.

Regarding training with HRV data only, Silveira focused on time and frequency HRV features, as the ones present in table 5.1. The inclusion of these variables should be considered carefully, as some frequency and non-linear HRV indexes are not valid for short intervals as the ones used in this and Silveira's work, that is 2 minute time windows.

Table 5.1: HRV features extracted in Silveira's work

Time features	Frequency features
HR	HF
SDNN	LF
SDSD	VLF
RMSSD	TP
NN50	nuHF
pNN50	nuLF
NN20	nuVLF
pNN20	LF/HF

From Silveira's work it is expected that the SVM and GBT are the best performing models, although her findings point that this means only 60% accuracy, with the minority class (drowsy) being classified almost completely at random. Also, the neural network is expected to present very poor results, most likely because the data available is not enough to avoid fitting of noise present in it.

5.1 Models training

To establish a baseline, the models were firstly trained using the entire dataset with a 70% split between training and test. To ensure generalization of the results, each model was trained and tested ten times with independent splits. Before training, ten-fold cross validation was used to tune each models' hyper-parameters, that that all experiments here presented reflect the effect of data and model capability on the models' performance. Features were also normalized using z-score transformation. Four metrics of performance are calculated after the tests, being those the accuracies of the model, and for each of the classes, as well as Matthews Correlation Coefficient (MCC). While the first is used to compare results with those of Silveira's, class accuracy is used to understand the distribution of misclassification, and finally Matthews is used to correctly appreciate the generalization capability of the model.

In this first training scenario, model accuracy behaved as expected, SVM, one class SVM (ocSVM) and GBT had accuracies between 60% and 70%, and the neural network performed poorly with only 41% accuracy. However, it is to note that this results are misleading: the classes in this dataset are imbalanced, with two thirds of the data points belonging to the alert class. Looking at the class performances, its visible that indeed the models are classifying almost every sample as alert, being the accuracy for drowsy samples very low in those first three models. This imbalance is well identified by MCC, that is close to zero, indicating that the models did not learn any structure in the data. Regarding the neural network, it presented a close to completely random behaviour, again indicating that no structure of data was learnt. However, for the ANN it seems that class imbalance is not to blame, but that the network is not able to adjust with the available data.

To test if the uneven distribution of data through classes was the real major factor for these poor performances, class weights were defined, so that the error of misclassifying a sample of the minority class matters enough to sacrifice the correct labeling of some of the majority class samples. This way each class weigh was defined as the proportion of data that the other class corresponded to. As the ocSVM does only sees one class during training, there is no class weight definition for it.

Figure 5.2 shows the results from this second test. As pointed before, ocSVM is unaltered by this class balancing. Both SVM and GBT increase their performance regarding the classification of drowsy samples, but at the cost of the other class, presenting an only slightly improved MCC. The neural network continues to behave randomly, corroborating the hypothesis that it has not enough data to learn this classification task.

This low performances propelled the need to investigate alternative forms of training. One of which was to train individual models. Using the fact that the specific cardiac dynamic one individual has is quite specific, as it depends on several factors: age, weight, physical fitness, *etc.*. Because of this, the regions of HRV that a person occupies when feeling alert and transitioning to drowsy will be personalized, thus making it virtually impossible to find a single separation rule for an heterogeneous population. Previous

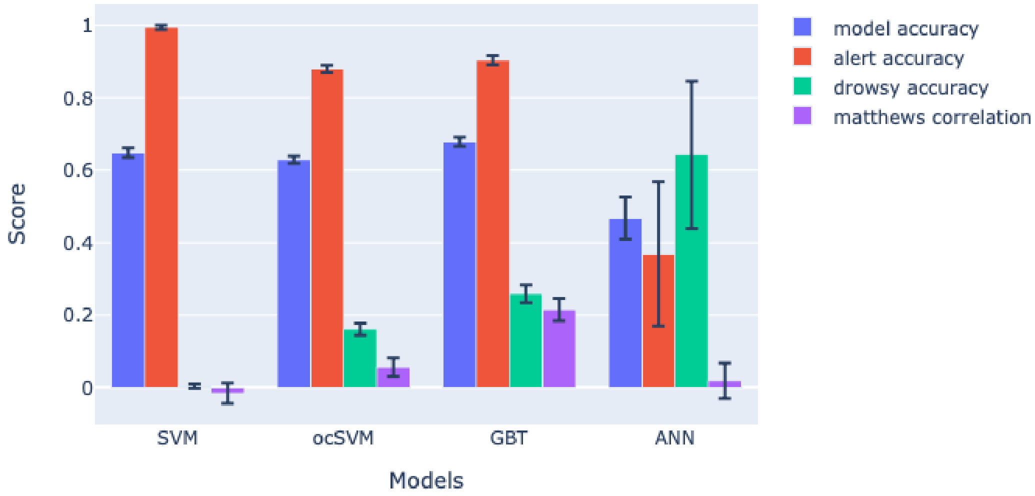


Figure 5.1: Models performance with initial training strategy, data imbalance leads models to perform poorly.

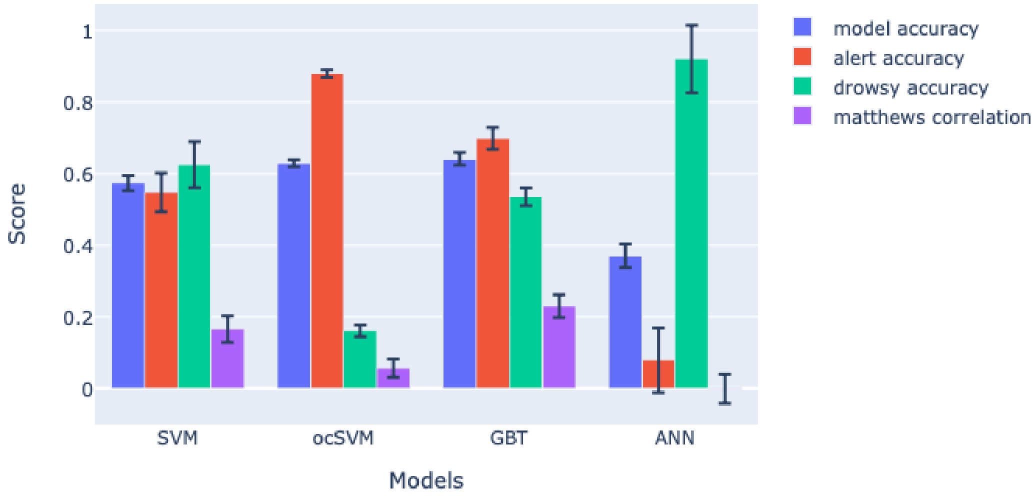


Figure 5.2: Models performance with class weights, models continue to perform poorly, leading to the proposition that other factors contribute to the difficulty of this classification task.

works have already pointed that this individualized strategy results in improved results [129].

5.1.1 Individual training

Here models were trained using the original procedure and hyper parameter, but different instances of each model were trained, each with the data of only one subject. Tables 5.2 report the metrics results on each model-individual pair. It is visible that SVM and GBT benefited from this strategy, having generally significantly improved performances. This table confirms also that the complexity of ANN is the sole cause for its poor results, achieving a mean -0.02 ± 0.04 MCC across subjects. The results obtained by training ocSVM show that assuming drowsiness to behave as an outlier does not lead to positive results, with an overall MCC of -0.17 ± 0.04. This may be cause by the scarcity of data, leading this model to define an outlier threshold that does not generalize well, misclassifying most of the test dataset.

From the four proposed models, only two seem to perform satisfactorily in this classification task, the SVM with 0.42 ± 0.03 MCC and GBT with 0.40 ± 0.04.

Looking in more detail to the different subjects, its possible to visualize that while most subjects provide data that fits high performing models, with MCC above 0.5, some individuals present values close to zero, namely FP07, FP15 and FP19.

Table 5.2: Matthews correlation coefficient for each model-subject pair. Best performing model for each individual is marked with bold.

	SVM	**ocSVM**	**GBT**	**ANN**
FP01	**0.69±0.15**	-0.36±0.25	0.62±0.19	-0.03±0.25
FP02	**0.66±0.11**	-0.40±0.16	0.46±0.11	-0.01±0.14
FP03	**0.94±0.04**	-0.55±0.17	0.77±0.06	0.00±0.00
FP07	-0.02±0.03	0.04±0.18	**0.04±0.20**	0.00±0.08
FP08	0.34±0.13	-0.19±0.16	**0.42±0.12**	0.00±0.00
FP09	**0.43±0.12**	0.01±0.11	0.38±0.16	-0.02±0.21
FP10	**0.65±0.10**	0.21±0.23	0.49±0.11	0.00±0.00
FP11	**0.53±0.11**	-0.15±0.15	0.43±0.17	0.10±0.28
FP12	**0.64±0.11**	-0.12±0.17	0.27±0.22	-0.04±0.18
FP13	0.29±0.18	-0.20±0.18	**0.38±0.22**	0.09±0.15
FP14	**0.51±0.20**	-0.22±0.16	0.31±0.19	0.00±0.00
FP15	-0.01±0.14	-0.18±0.15	**0.11±0.09**	0.00±0.00
FP16	-0.02±0.09	0.02±0.03	**0.54±0.25**	-0.18±0.21
FP17	**0.56±0.08**	-0.28±0.11	0.42±0.21	0.00±0.00
FP18	**0.93±0.08**	-0.19±0.21	0.67±0.21	-0.17±0.44
FP19	**0.06±0.18**	-0.09±0.15	-0.02±0.25	0.00±0.00
FP20	0.52±0.13	-0.22±0.18	**0.55±0.20**	0.00±0.00

To evaluate what was causing these outliers, t-SNE plots were created for each subject, as a visual measure of the class separability their dataset provided. Their KSS annotations were also plotted, to try and base the separability assessment on the quality of those labels.

Figure 5.3 serves as a comparison point, as it represents the data separability of subject FP03, one of the best performing ones. There is visible a clear difference on the driver's reported state between the day and night session. And that separation of states results in separable classes as seen in the t-SNE plot.

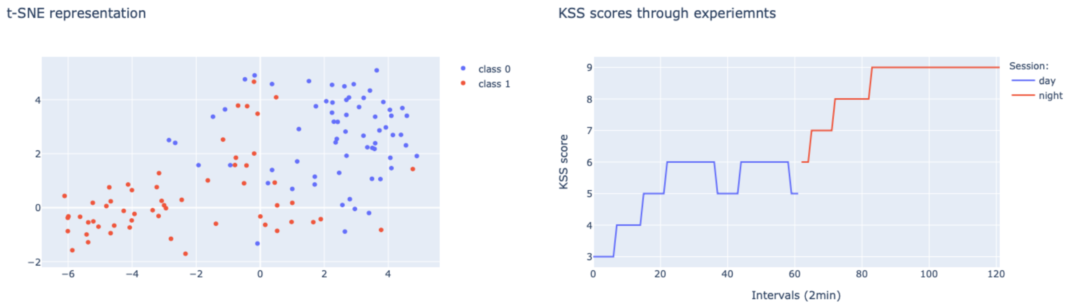


Figure 5.3: t-SNE plot and KSS distribution for subject FP03. The excellent performance of SVM and GBT models for this subject are a result of the quality of KSS scores, with a clear difference between day and night driver state, which is reflected in a t-SNE plot with a clear separation between the two classes.

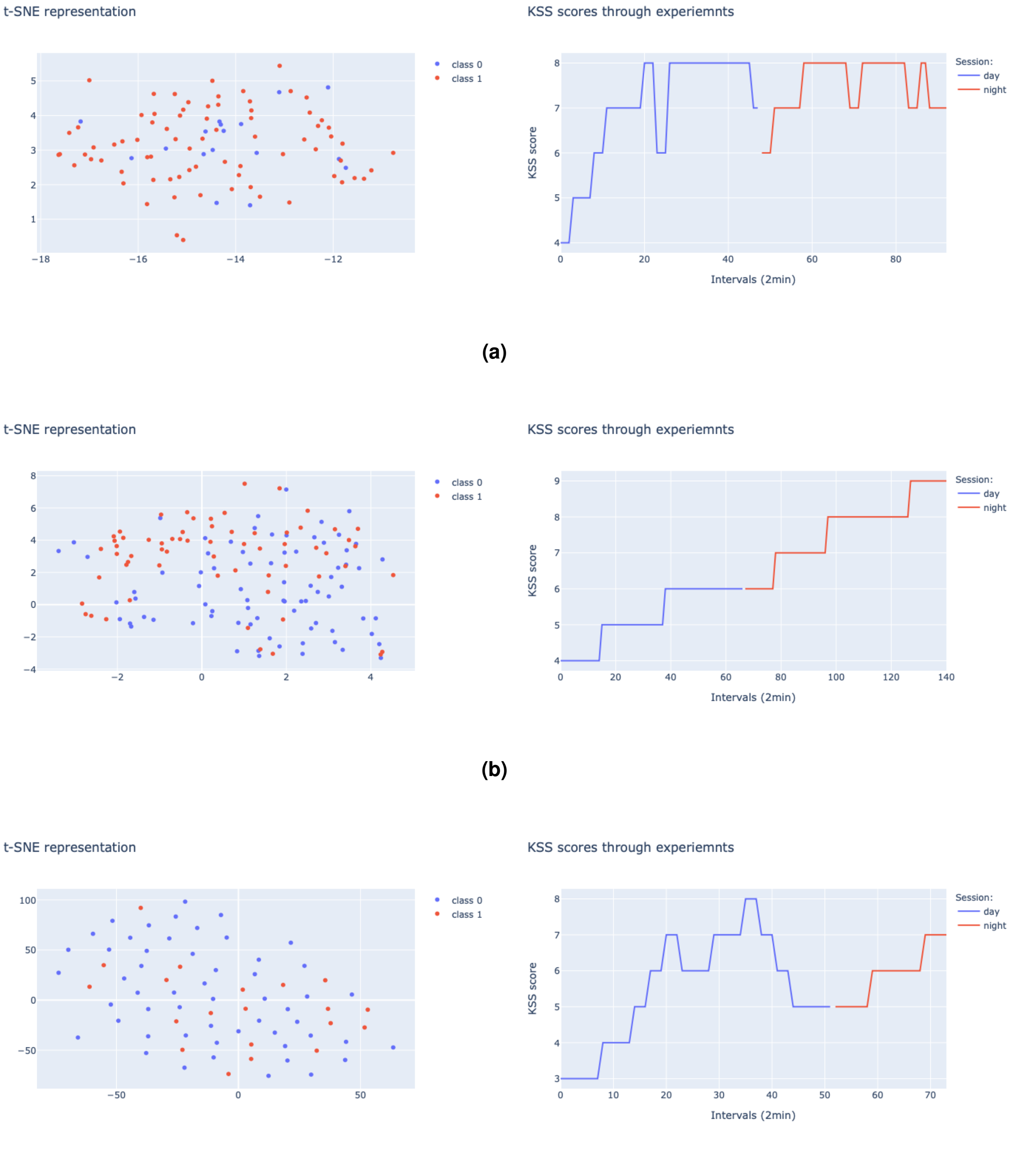
Figure 5.4 represents different cases where classification performances were extraordinarily low: in 5.4(a) and 5.4(c), a large majority of points belongs to a single class, alert and drowsy respectively. The failure of these models can also be related to the degree of variation the KSS reports present, suggesting some difficulty by these individuals to correctly assess their state, which explains why the minority class points are completely mixed with the majority ones. 5.4(b) presents another problem that lowers the performance of these models, but not to the extent of the previous cases, that is the continuous nature of the drowsiness process, there are preferential concentrations of points in different sides, but a substantial overlap between clusters lowers the classification performance.

This finding is related to the main flaws pointed at KSS, in that it is highly dependent on the subject's capability to correctly evaluate their state, and to understand what the different levels represent. Failing to produce accurate KSS jeopardizes all future classification that depends on that data.

In the absence of an objective drowsiness scale, as JDS, is important to guarantee that subjects being measured to produce a dataset for drowsiness classification fully understand the scale, and can distinguish drowsiness from other feelings, as boredom. Also, measures to induce alertness and drowsiness must be studied and implemented, in order to minimize class overlap due to the continuity of drowsiness process.

Eliminating the individuals that presented these flaws (FP07, FP08, FP13, FP15, FP16 and FP19), improves the average performance of SVM and GBT models, respectively to this way better expressing their fitness for this task, given acceptable data the achieve, respectively, 0.64 ± 0.04 and 0.49 ± 0.05 MCC scores. The significant increase in average performance of SVM sets it as the fittest model for this task of personalized classification of drowsiness using HRV features.

To confirm that the outlier subjects were not causing the poor results of the original training, but



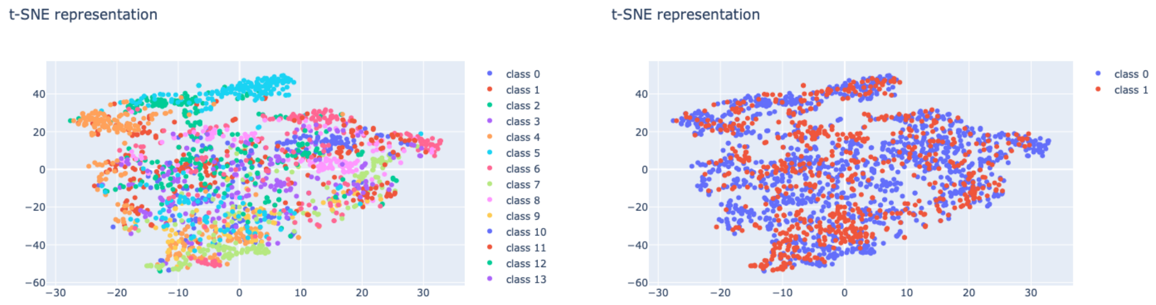
(a)

(b)

(c)

Figure 5.4: t-SNE and KSS representations of poorly performing individuals, respectively (a) FP07, (b) and (c) FP19, representing three major problems that KSS annotated data can have for the drowsiness classification task: class imbalance, annotation error and state continuity.

that the individualization of models was actually the key factor, global models were retrained using only data from the well performing individuals. The removal of those outlier datasets has no impact, as all models resulted in MCC close to zero. To visualize the importance of model customization, t-SNE representations of the well performing dataset were created and presented in figure 5.5, coloring points according to the subject they come from, or to the class they belong to. There it becomes clear the importance of individual models, as each subject creates its own cluster of HRV data, which results in a complete mixture of alert and drowsy data points in a general dataset, as their separation exists only in an intra-subject scope.



(a)

(b)

Figure 5.5: t-SNE representation of the dataset formed by all the well separated individuals data. While both plots distribute the same data, (a) colors each point according to the subject the point comes from, and (b) colors the points according to the class (0=alert, 1=drowsy) they belong to.

5.2 Model Architecture and training procedure

Having defined SVM as the best model for this task, it became important to further tune the model in order to maximize its performance after fitting individual data. To do so, features were revised, using the *a priori* knowledge about the validity of certain HRV indexes in the context of usHRV, and selected, using unsupervised methods. Model hyper-parameters were also fine tuned, performing an grid search to find the parameters that produce the best average score across the data set population. Finally, class balancing strategies were tested to increase the generalization capability of the fitted models. At the end of this process, a set of features, hyper-parameters and a class balancing strategy were defined to deploy a model architecture that only needs training data to start being applied to new drivers.

5.2.1 Feature Selection

Until this point, the features used in the model were those listed in table 5.1. However, ensuring an optimal set of features can increase the performance of the model. Before implementing any algorithm on the feature set to select the best features, a revision on each features validity in respect to their application on usHRV, as the two minute analysis intervals require, is conducted.

While there are no constraints on which time-domain features to use in these short time windows, frequency ones loose meaning when the analysed window doesn't allow the needed resolution. This is the case when trying to use VLF and normalized VLF with only two minutes of signal, as the lower bound of this frequency range is $0.004Hz$, at least five minutes of signal would be needed to have enough frequency resolution to calculate these indexes. For this reason, these two features were discarded.

Initially no nonlinear-domain features were used, and, although most chaos indexes need longer period of time to separate chaotic from random behaviour, two short-term related features can be added to the dataset. They are the first *alpha* component of DFA and the Poincaré SD_2, which can be calculated meaningfully for sets of tens of points, and two minutes of signal contain normally 120 IBIs [130].

Changing the features as described before changes the model's average performance from 0.64 ± 0.04 to 0.61 ± 0.04. Even though it is a reduction, given that the difference is smaller than the uncertainty, and that the new set of features is more in line with the prior knowledge of feature validity for usHRV, the new set of features is maintained.

Finally, to ensure that no useless feature was being used, unsupervised feature selection was implemented, using MAD as a relevance measure, a and correlation coefficient as the similarity coefficient. As the number of features is already relatively low, only 16, the thresholds on the selection algorithm were soft, allowing enough features to have 98% cumulative relevance, and allowing every feature with less than 0.95 correlation coefficient to be accepted in the final feature set. Of all features, the only that seem be disposable was LF, and, retraining the individual models with the feature set that excludes this one slightly increases the average performance from 0.61 ± 0.04 to 0.62 ± 0.04.

5.2.2 Hyper-parameter tuning

To refine the SVM parameters optimally for this classification task and datatype, a grid search over the model's parameters was performed, evaluating the parameter sets using the average MCC as the performance metric. This way a set of parameters is chosen that optimizes the performance of this model trained for any subject, and not for a specific subject in detriment of another.

Four hyper-parameters were selected, kernel type, that defines the function basis of the kernel used for the learning task, penalty parameter C, that controls how strict or loose is the misclassification permission in order to manage its trade-off with margin maximization, the parameter gamma, that is a parameter of polynomial and Radial Basis Function (RBF) kernels, and the degree of the polynomial kernel.

Table 5.3 lists the values tested for each parameter considered in the grid search.

Table 5.3: Parameter space searched evaluated through grid search.

Parameter	Values
Kernel	'linear', 'poly', 'rbf', 'sigmoid'
C	0.1 - 2, in increments of 0.2
gamma	'auto', 'scale'
degree	2, 3, 4, 5

Running this search returns that the best set of hyper-parameters is the linear kernel with an regularization parameter C of 0.3, increasing the average MCC by only two thousandths from the original set of hyper parameters, that were linear kernel with C equal to 0.5. The small increase in the performance means that its rounded value continues to be 0.62 ± 0.04, as the original parameters were already very similar to the optimal ones.

5.2.3 Class balancing

To finalize the improvement process of this learning task, a simple class balancing algorithm was implemented, SMOTE, to help improve the models performance when detecting the minority class. This method creates new data points for the minority class, by placing them in between close existing points of that class, this way guaranteeing the maintenance of that classes dispersion, but better populating its cluster, so that the machine learning algorithm has to define more accurate boundaries between classes in order to minimize its cost function. By implementing this method performance increases by a few thousandths again, setting its final value as 0.62 ± 0.03.

5.2.4 Final model

The process previously described defined the model to use and its training strategy: a SVM model with linear kernel and regularization parameter C 0.3. It should take 15 features as inputs, those listed in table 5.4 and output a label stating if a certain time period corresponds to alert or drowsy. One model should be fitted for each individual, with the training set balanced using SMOTE and data further standardized.

Table 5.4: Features to use in the final drowsiness classifier.

Time-domain	Frequency-domain	Nonlinear-domain
HR	nuHF	DFA α_1
NN20	TP	SD2
NN50	LF/HF	
pNN20	HF	
SDNN	nuLF	
RMSSD		
pNN50		
SDSD		

6

Implementation of drowsiness detector on driving simulator

Contents

Transitioning from traditional chest ECG to the peripheral cardiac signals' based HRV demands a comparison of results obtained with each signal. To do so, a database containing simultaneous recordings of chest ECG, hands ECG and wrist PPG, in a driving context, was needed to produce the data that would provide answers on whether such transition was possible.

In coordination with other projects being developed at CardioID, an experimental setup was designed, making use of the AUTOMOTIVE: AUTOmatic multiMOdal drowsiness detecTIon for smart VEhicles [131] simulator and the three physiological data sources discussed in chapter 4 (CardioWheel, PulseOn wrist band and Movesense).

6.1 AUTOMOTIVE - the simulator

The AUTOMOTIVE project aimed to be a platform able to investigate sleepiness detection through an array of different sources of information, namely steering wheel dynamics, eye and gaze detection and HRV. It incorporates these inputs with a virtual environment, that emulates two cities connected by a high way.

While the cities have their architecture defined, the highway is built in a procedural form, which allows its customization in accordance with the needs of any experiment, in terms of road length, and ensures that every simulation run has a slightly different sequence of turns in the highway. The environment can also be populated with Non-Player Character (NPC) cars and pedestrians, making it more realistic and more cognitively demanding for the driver.

The subject controls the virtual car by means of a set of pedals, for acceleration and braking/reverse, and a steering wheel equipped with the CardioWheel.

For its multi-modal purpose, the simulator also integrates a camera capable of detecting facial features and estimating gaze direction and detecting blinks, which was not used in the experiments conducted in this thesis.

Finally, to provide annotations regarding the alertness state of the driver, the system can prompt a KSS scale where a value is selected, saving these annotations alongside with the IBIs from CardioWheel.

Data produced by the simulator and CardioWheel is communicated to a SQL database, identifying the session in which it was collected, the driver by a Universally Unique Identifier (UUID) and the timestamps of each datapoint.

6.2 Experimental design

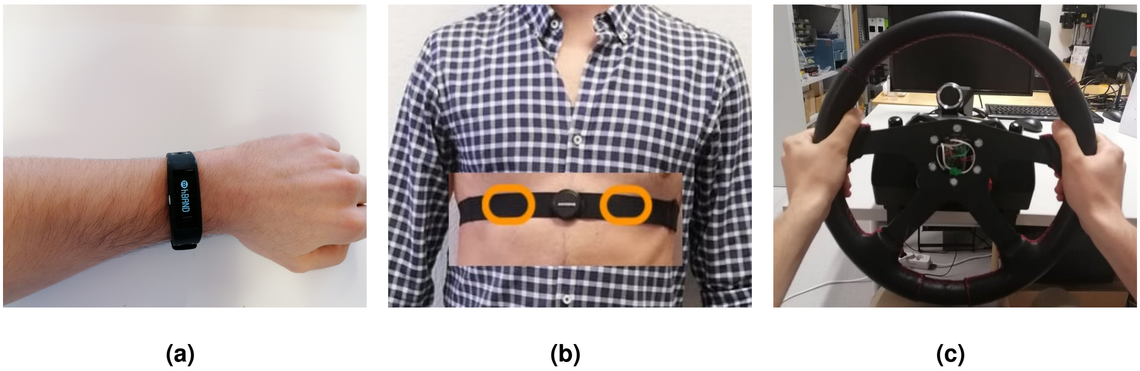
The data collection consisted in double sessions of 30 minutes, for each tested individual.

Before beginning the experiment, subjects are briefed on the objective of the project, and have the KSS explained to them, as a message prompts every five minutes asking for their sleepiness evaluation. They are asked to state their self assessed state every time the message appears, without any intervention by the experiment supervisor, which would only store the stated value. The scale used was the version adapted for Portuguese, retrieved from [132].

The drivers are also asked to maintain the car in the rightmost lane at all times, except if overtaking a NPC car, and to keep an average speed of 60km/h throughout the experiment. This is done to ensure constant need of engagement by the drivers, while the activities are not demanding enough to prevent drowsiness and fatigue to set in.

The number of sessions each subject attends to, two, is used to have sessions where alert and drowsy states are promoted. Alertness is stimulated by having the session in the middle of the morning after a good night sleep (>8 hours), and by setting the simulator environment lighting on a sunny day mode. Drowsiness is boosted by having subjects being measured at the end of the day, with the past 24 hours without any coffee, tea or energetic drinks, or after a night with short sleep (<6 hours). In these sessions the simulator would be set to night mode, to further promote an environment prone to sleepiness and fatigue.

All 30 minute sessions consist of a highway connecting two cities, with approximately 30km. The driver starts by crossing an intersection and immediately entering that highway. During the first 5km, NPC cars are generated every 500 meters, with constant velocities that range from 50 to 70km/h, forcing the driver to adopt dynamic strategies on whether to overtake or follow these agents, while respecting the traffic rules. After those 5km, the driver only has to maintain its speed and keep the car inside its lane.



(a) **(b)** **(c)**

Figure 6.1: Recommended placement of sensors: (a) wrist band, fastened bellow the wrist bone; (b) chest band, just bellow the pectoral muscles and with each contact (marked in orange) on one side of the chest; (a) CardioWheel, hands symmetrically placed at a middle height.

6.3 Results

In total, 13 volunteers (2 female) with ages 33.42 ± 10.90 were recruited. From those, only 10 completed both sessions, resulting in 23 simulation trials. Three sessions were eliminated from analysis for belonging to incomplete data collection of three different individuals, and additional four sessions were discarded because simulator data was corrupted for two sessions of different subjects, rendering the four sessions associated with them unusable. Finally, of the eight subjects to be analysed, four reported the same state (either alert or drowsy) for the totality of their sessions, so they had to be discarded as well from analysis, due to lack of different class samples to train and test the models.

The final dataset consisted of eight sessions, two for each of four individuals (1 female) with ages 29.75 ± 12.85. For each individual, Movesense's ECG and PulseOn's PPG signals were processed as described in chapter 4, and had their IBIs extracted. CardioWheel already provided IBI values, so those were used directly. Using the stored KSS values and their timestamps, the annotation values were upsampled of to 1 KSS value per minute with linear assigning each sample the value of the nearest original one, and the resulting intervals were used to select the IBIs values to evaluate for each signal in two minute intervals, with 50% overlap. Overlapping was used to increase the amount of available data, as each participant collected around 60 minutes of signal, using two-minute windows without it would half the number of samples. The HRV features listed in table 5.4 were calculated for each interval and device, creating three sets of features for each signal source.

Finally, a system to train and test models with the architecture defined in the previous chapter was implemented. For each individual, a table with features and KSS values from both sessions, and for each device was loaded, and, in 10 different runs, each dataset would be separated into train and test portions with 30% assigned to testing. Three models would be trained using the training sets of each device. The models of each device were then tested using the Movesense model on all three sources test data, and the respective test sets with the PulseOn and CardioWheel models. The rational of this was to have a comparison of performance between Movesense model on the defined ground truth data and the peripheral signals data, but also to obtain a measure of information contained in the wearables, having the performance of the model trained in the same wearable data. The metrics presented in table 6.1 are mean and standard deviation of MCC across the 10 runs.

Two very different patterns appear in these results, on one hand, subjects 152 and 128 have very high MCC values for the models trained and tested with only one data source, always above 0.60 and hold very decent performances when the Movesense model is used on the other datasets, being remarkable the fact that for individual 152, PulseOn data is classified as correctly by the Movesense model as it is by its own model. On the other hand, participants 170 and 147 present lower performances in their one-device only models, and fail completely to classify samples from the peripheral devices using the Movesense model.

Table 6.1: Results for signal source performance comparison, for each subject, columns correspond to the origin of test data, and rows to training data.

Subject 152

	Movesense	CardioWheel	PulseOn
Movesense	0.81±0.10	0.46±0.26	0.61±0.20
CardioWheel	–	0.77±0.11	–
PulseOn	–	–	0.61±0.19

Subject 128

	Movesense	CardioWheel	PulseOn
Movesense	0.70±0.13	0.34±0.20	0.54±0.21
CardioWheel	–	0.62±0.13	–
PulseOn	–	–	0.70±0.14

Subject 170

	Movesense	CardioWheel	PulseOn
Movesense	0.41±0.30	-0.06±0.15	-0.03±0.21
CardioWheel	–	0.41±0.18	–
PulseOn	–	–	0.14±0.22

Subject 147

	Movesense	CardioWheel	PulseOn
Movesense	0.47±0.48	0.00±0.00	-0.09±0.18
CardioWheel	–	0.27±0.42	–
PulseOn	–	–	0.40±0.49

Looking further into the original sample distribution in all four individuals, one finds that while subjects 152 and 128 have fairly good class distributions, with 18/32 and 17/41 drowsy/alert sample ratio respectively, the other two present more skewed class balances, with 9/48 and 4/45 corresponding ratios. The hypothesis is that such level of imbalance results from phenomena described in chapter 5, related with the frailties of KSS annotations. In fact, both this low performing individuals report a maximum level of KSS of seven (sleepy (not fighting sleep)), and only at the end of of a session where initial levels were as low as two (very alert/alert).

With this analysis, there is an clue that a drowsiness detection system based on HRV and agnostic to the data source is possible, considering that the two best performing individuals have good performances even when tested data originates from a different device from that with which the model was trained. Granted that the reduced population in which these results were obtained demands future work to validate such claim, it indicates a possibility that must be considered.

To support statistically that the classifications of the different signals tend to be equivalent, the best model trained with Movesense data is selected to classify the entirety of the three datasets. That creates three binomial distributions in which McNemar's test can be used to evaluate their similarity.

McNemar's test is the statistical test to use when determining the difference between two dependent binary populations. As stated by McNemar [133], the test is to be applied in situations like the change in responses on a questionaire by the same population after an experiment, or the difference in responses to two different questions by the same population, in order to evaluate if the experiment changed the population answers, or if the questions had significantly distinct difficulty levels. In this particular case, it is intended to evaluate if the classification of the same physiological quantity, HRV, becomes significantly different when a different base signal is used to observe said quantity. McNeamr's test statistic is calculated by equation 6.1, where b is the number of samples where one population had true predictions and the other false ones and c is the number of samples a population had false predictions and the other true ones. The statistic is approximated by a χ^2 distribution with one degree of freedom [133]. However, when the number of disagreeing data pairs is small, as it happens to be in all the cases studied here, it is best to compare it with a binomial distribution with size equal to the number of disagreeing pairs and $\theta = 0.5$ [134], which was used to calculate the p-values in table 6.2.

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{6.1}$$

In all cases, the number of different classifications between two datasets is small, so the p-value of this test can be calculated using the exact binomial test, for which the results, defining null hypothesis rejection at p-value < 0.10, are presented in table 6.2. The null hypothesis, that the binomial distributions are equivalent, is accepted for all pairs.

Table 6.2: McNemar's test results regarding similarity of classifications from different sources.

Subject	Pair	χ^2	p-value	result
152	Movesense+CardioWheel	1.00	0.508	Equivalent
	Movesense+PulseOn	3.57	0.125	Equivalent
128	Movesense+CardioWheel	0.50	0.727	Equivalent
	Movesense+PulseOn	1.29	0.453	Equivalent

7

Conclusion and Future Work

This thesis studied the feasibility and technical requirements of producing a peripheral cardiac signal based drowsiness detection system. Three main dimensions of this problem were approached and answered: how to collect these peripheral signals and convert them into streams of IBIs values, how to use those values in drowsiness detection and whether such system could be agnostic to the original signal measured.

The first question was answered by introducing three different types of wearable devices capable of collecting cardiac rhythm information: the chest strap Movesense, the capacitive steering wheel CardioWheel, and the wrist PPG sensor PulseOn. While the ECG based devices provided built-in filtering that allowed direct detection of R peaks and subsequently IBIs, the PPG sensor suffered from sensitivity to external conditions, such as perceived ambient illumination from both light and hand position changes. This created sudden offset changes in the signal that needed a special filter to eliminate, that could not depend on frequency filtering due to the step nature of those artifacts. Instead, an online filter that mimics recursive moving average removal was created. By applying such filter to mimic a window of one second, resulting signals would maintain only the oscillatory component, where cardiac rhythm is encoded. Additionally, an adaptive threshold peak detection algorithm was implemented to locate the peaks of PPG signal. the algorithm used also a refractory period of 0.6 times the length of the last detected inter-beat interval to avoid false peak detection, and reset its threshold after 1.5 times the last detected interval passes without a new peak detection. This created a detection system robust against changes in pulse amplitude, false peaks created by sudden movement and interruptions in the signal pulses. For the other devices, while CardioWheel directly provided the IBIs itself calculated, Pan-Tompkins algorithm was used to detect R peaks in Movesense signals, correcting the peak locations by selecting the maximum value in a 0.4 seconds window centered in the initial estimates.

While the peak detection methods used in the different signals proved capable of identifying the peaks present in them, moments of poorer contact between the individual and the devices lead to missing peaks and added artifacts that corrupted some of the intervals collected. While normal procedures to treat such outliers consist of simply eliminating them, the ultra short nature of the analysis time windows used, 2 minutes, required a more conservative approach. Hence a IBI corrector system was created, evaluating the ratio between consecutive IBIs to detect both missing peaks and false detections to accordingly estimate the location of the non-detected peak and divide the longer IBI or two join two shorter intervals into the true IBI. By testing this system on artificially corrupted segments of visually validated ECG, it was shown that the system is capable of reconstructing the sequence of IBIs from signals corrupted with 10% missed detections and additional 10% false peaks with less than 7.5 milliseconds of mean absolute deviation from the true signal. The system was tested to the limit of having 40% of the signal values corrupted, and still managed to retrieve a stream of values with an MAD of 38.03ms, which, while very unlikely that such a large portion of the signal produces faulty IBI values, its still a smaller temporal deviation than the uncertainty in IBI determination on a 25Hz signal as the PPG is. This system is relevant to ensure that all collected information is used to calculate the HRV as confidently as possible, but the author leaves also the suggestion of its usage on analysis of longer term HRV, as it maintains the true succession pairs used in non-linear analysis as the Poincaré plot and the traditional sample elimination does not.

The second question, how to use the IBI values to detect drowsiness, was answered by searching the best subset of HRV features and the best model architecture to do so. An initial set of time and frequency domain features was used to compare four decision models, SVM, ocSVM, GBT and a ANN. In the process of testing which model performed best, the author realized that a general models, this is, a model trained to classify drowsiness in any arbitrary individual was performing poorly, independently of model architecture. This lead to the investigation of personalized models, which showed great improvements for part of the population. The individuals that continued to perform badly showed upon further analysis of their data that the limitations of the experiment and sleepiness scale used for the database, SleepEye, brought:

- Unbalance in classes, while the experiment was designed to have both alert and sleep deprived driving sessions, not all participants managed to provide enough KSS ratings associated with being sleepy for the model to properly learn the separation boundary between the two classes, even with class balancing methods applied.
- Imprecise self rating of their own state, being KSS a subjective drowsiness scale, the confidence in the annotations is proportional to the capability each individual has to self assess its state and correctly understand the levels of the scale. By looking into some of the ratings provided by subjects in this dataset, consecutive values with high ranges of variation raised the suspicion that some individuals were not accurately reporting their KSS level.
- The fundamentally continuous nature of drowsiness, as it is not a biological switch, where people would be either fully alert or fully drowsy, it is a process that sets in continuously, which makes the definition of a dangerous drowsiness level a rather arbitrary process, blurring the class separation in this problem. It was observed that the best results were obtained by individuals that reported both very low KSS values (¡4) and high ones (¿7), while those that concentrated ratings in values between 5 and 7 had the poorest performances.

By evaluating only the population whose annotations showed a good understanding of the scale, trustworthy self report and balanced experience of both alert and drowsy states, the models trained and tested for each of the 12 selected individuals attained a mean performance across them of 0.64 ± 0.04 and 0.49 ± 0.05 MCC for SVM and GBT respectively, while the other two models continued to perform poorly, thus being discarded. At last, to confirm that the poorly performing individuals were not the cause why the general model failed, a new general model was trained and tested with data from the 12 selected ones. This new model failed, and it was shown that the real reason for that was that each individual forms its own cluster in feature space, and while a frontier can be defined between the alert and drowsy data of one individual, the displacement of the various subjective clusters makes it impossible to determine a single common boundary.

From here, the SVM model was selected as the best fitted to classify personalized state of drowsiness. Features used in the classification were revised, eliminating VLF because it did not hold significance when calculated in a short time window as 2 minutes, and two non-linear features were added: first α component of DFA and Pointcaré SD_2. Unsupervised feature selection using MAD as the relevance metric was applied, which resulted in the elimination of LF feature. And Finally, hyper-parameters were fine tuned, defining an SVM with linear kernel and C parameter 0.3 as the best architecture for drowsiness detection, which attained a mean performance of 0.62 ± 0.03 MCC, which indicates a strong correlation.

This answers the question on whether IBI values can ultimately be used to detect drowsiness, but all these models were tested and trained with data collected through a chest ECG, and a final question must be analysed, can the same model detect drowsiness, but from IBIs measured from a peripheral signal?

The experiment conducted in this thesis aimed to answer that. By applying the tools developed in the rest of the work, the simultaneously collected signals (chest ECG, hands ECG and wrist PPG) were converted into IBIs, and ultimately, HRV features were calculated for every two minute window in each of the signals. Unfortunately, from the 13 volunteers recruited, only two managed to survive the selection criteria applied to the SleepEye dataset. Three individuals had only one session, two individuals had missing data in one of their sessions, and four had reported the same state (either alert or drowsy) throughout the two sessions, not providing the two classes needed for training and testing the models. Additionally, two individuals had very unbalanced class distributions, and suffered from the same limitations found in the SleepEye data annotations, KSS values all bellow 8, and very few minutes reporting a 7 drowsy state.

The two individuals that survived the selection criteria produced models with very good MCC scores when trained and tested with data from the same device, with all devices. Those scores ranged from 0.62 to 0.81. And, to answer the final question, the model trained with data from the Movesense device, remained well performing when applied to data from the peripheral cardiac signals, ranging scores from 0.34 to 0.61, and with the high note of the performance of classifying PulseOn data with the model trained with its own data or the Movesense one is the same. Additionally, McNemar's test was used to compare the classifications of the entire dataset of each device with the Movesense trained model, for each individual, and it showed that all crossed classifications were statistically equivalent to the base classification, Movesense train on Movesense data.

This results indicates that the system this thesis proposes is very possibly feasible, and well performing.

However, future work has to be developed to affirm this with certainty, firstly to compensate the limited size of the analysed population. A new study has to be conducted to evaluate if the findings of this work hold. The reduced number of individuals here analysed is pointed as the main limitation of this work, however the recruitment of volunteers during this pandemic time was mostly nonexistent, and the recruits consisted of the CardioID team and two professors involved in the development of the Simulator. The time window that aligned the readiness of the simulator with the ease in lockdown measures, as well as the calendars of each of the participants, unfortunately didn't allow the retake of some of the sessions to be able to add more individuals to the final analysis.

One other interesting point to develop in the future is the fact that while individualized models proved to be the possible way to detect drowsiness, training each model for every new user of this system is not doable in a market perspective. However, it is hypothesised that a limited set of individual models can be representative of the possible ranges of HRV for a general population. By finding such set and combine them in a voting scheme or other ensemble classification framework a general and ready to apply drowsy detection system based on HRV can be created.

Bibliography

[1] "Road traffic injuries," https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, accessed: 2020-09-24.

[2] N. H. T. S. Administration, "2016 fatal motor vehicle crashes: Overview," NHTSA's National Center for Statistics and Analysis, 1200 New Jersey Avenue, SE Washington, DC United States 20590, Tech. Rep., 10 2017.

[3] A. A. Alian and K. H. Shelley, "Photoplethysmography," *Best Practice & Research Clinical Anaesthesiology*, vol. 28, no. 4, pp. 395–406, 2014, hemodynamic Monitoring Devices. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1521689614000755

[4] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals," *Curr Cardiol Rev*, vol. 8, no. 1, pp. 14–25, Feb 2012, [PubMed Central:PMC3394104] [DOI:10.2174/157340312801215782] [PubMed:20702919].

[5] C. Park, H. Shin, and B. Lee, "Blockwise PPG Enhancement Based on Time-Variant Zero-Phase Harmonic Notch Filtering," *Sensors (Basel)*, vol. 17, no. 4, Apr 2017.

[6] W. Waugh, J. Allen, J. Wightman, A. Sims, and T. Beale, "Novel signal noise reduction method through cluster analysis, applied to photoplethysmography," *Special edition: Computational and Mathematical Methods in Medicine*, vol. In Press, 01 2018.

[7] E. Sabeti, N. Reamaroon, M. Mathis, J. Gryak, M. Sjoding, and K. Najarian, "Signal quality measure for pulsatile physiological signals using morphological features: Applications in reliability measure for pulse oximetry," *Informatics in Medicine Unlocked*, vol. 16, p. 100222, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352914819301856

[8] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, "An optimal filter for short photoplethysmogram signals," *Scientific Data*, vol. 5, no. 1, p. 180076, 2018. [Online]. Available: https://doi.org/10.1038/sdata.2018.76

[9] S. D.S., K. M. C., A. N., L. Janardhan, and R. H.S., “Case study on measurement of spo2 from ppg signals in the presence of motion artifact,” in *2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT)*, March 2017, pp. 318–323.

[10] Y. Ye, W. He, Y. Cheng, W. Huang, and Z. Zhang, “A robust random forest-based approach for heart rate monitoring using photoplethysmography signal contaminated by intense motion artifacts,” *Sensors (Basel, Switzerland)*, vol. 17, 02 2017.

[11] X. Sun, P. Yang, and Y. Zhang, “Assessment of photoplethysmogram signal quality using morphology integrated with temporal information approach,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2012, pp. 3456–3459.

[12] M. R. Ram, K. V. Madhav, E. H. Krishna, N. R. Komalla, and K. A. Reddy, “A novel approach for motion artifact reduction in ppg signals based on as-lms adaptive filter,” *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 5, pp. 1445–1457, May 2012.

[13] C. Wei, L. Sheng, G. Lihua, C. Yuquan, and P. Min, “Study on conditioning and feature extraction algorithm of photoplethysmography signal for physiological parameters detection,” in *2011 4th International Congress on Image and Signal Processing*, vol. 4, Oct 2011, pp. 2194–2197.

[14] W. Karlen, J. M. Ansermino, and G. Dumont, “Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2012, pp. 3131–3134.

[15] V. Jeyhani, S. Mahdiani, M. Peltokangas, and A. Vehkaoja, “Comparison of hrv parameters derived from photoplethysmography and electrocardiography signals,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 5952–5955, aUX=ase,”Jeyhani, Vala” AUX=elt,”Mahdiani, Shadi”.

[16] A. J. Pappano and W. Gil Wier, “2 - excitation: The cardiac action potential,” in *Cardiovascular Physiology (Tenth Edition)*, 10th ed., A. J. Pappano and W. Gil Wier, Eds. Philadelphia: Elsevier, 2013, pp. 11 – 30. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780323086974000022

[17] S. Vadrevu and M. S. Manikandan, “Use of zero-frequency resonator for automatically detecting systolic peaks of photoplethysmogram signal,” *Healthcare Technology Letters*, vol. 6, no. 3, pp. 53–58, 2019.

[18] T. Viet Thang and W.-Y. Chung, “A robust peak detection algorithm for photoplethysmographic waveforms in mobile devices,” *Journal of Medical Imaging and Health Informatics*, vol. 7, pp. 1617–1623, 11 2017.

[19] H. S. Shin, C. Lee, and M. Lee, "Adaptive threshold method for the peak detection of photoplethysmographic waveform," *Comput. Biol. Med.*, vol. 39, no. 12, pp. 1145–1152, Dec 2009.

[20] A. Choi and H. Shin, "Photoplethysmography sampling frequency: Pilot assessment of how low can we go to analyze pulse rate variability with reliability?" *Physiological Measurement*, vol. 38, 02 2017.

[21] S. Béres, L. Holczer, and L. Hejjel, "On the minimal adequate sampling frequency of the photoplethysmogram for pulse rate monitoring and heart rate variability analysis in mobile and wearable technology," *Measurement Science Review*, vol. 19, no. 5, pp. 232–240, 2019. [Online]. Available: https://content.sciendo.com/view/journals/msr/19/5/article-p232.xml

[22] H. J. Baek, J. Shin, G. Jin, and J. Cho, "Reliability of the parabola approximation method in heart rate variability analysis using low-sampling-rate photoplethysmography," *Journal of Medical Systems*, vol. 41, no. 12, p. 189, 2017. [Online]. Available: https://doi.org/10.1007/s10916-017-0842-0

[23] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Medical and Biological Engineering and Computing*, vol. 44, no. 12, pp. 1031–1051, 2006. [Online]. Available: https://doi.org/10.1007/s11517-006-0119-0

[24] R. Gordan, J. K. Gwathmey, and L. H. Xie, "Autonomic and endocrine control of cardiovascular function," *World J Cardiol*, vol. 7, no. 4, pp. 204–214, Apr 2015.

[25] D. Petković and Z. Cojbasic, "Adaptive neuro-fuzzy estimation of autonomic nervous system parameters effect on heart rate variability," *Neural Computing and Applications*, vol. 21, 11 2011.

[26] J. Sztajzel, "Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system," *Swiss Med Wkly*, vol. 134, no. 35-36, pp. 514–522, Sep 2004.

[27] M. Fernandes de Godoy, "Nonlinear analysis of heart rate variability: A comprehensive review," *Journal of Cardiology and Therapy*, vol. 3, pp. 528–533, 2016.

[28] K. Murakami and M. Yoshioka, "Pulse transit time variability on a range of heart rates between resting and elevated states," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, Oct 2015, pp. 1579–1582.

[29] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2006, pp. 4289–4294.

[30] N. Pinheiro, R. Couceiro, J. Henriques, J. Muehlsteff, I. Quintal, L. Goncalves, and P. Carvalho, "Can PPG be used for HRV analysis?" in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2016-Octob, 2016, pp. 2945–2949.

[31] B. Vescio, M. Salsone, A. Gambardella, and A. Quattrone, "Comparison between electrocardiographic and earlobe pulse photoplethysmographic detection for evaluating heart rate variability in healthy subjects in short- and long-term recordings," *Sensors (Switzerland)*, vol. 18, no. 3, 2018.

[32] K. Georgiou, A. V. Larentzakis, N. N. Khamis, G. I. Alsuhaibani, Y. A. Alaska, and E. J. Giallafos, "Can Wearable Devices Accurately Measure Heart Rate Variability? A Systematic Review," *Folia medica*, vol. 60, no. 1, pp. 7–20, 2018.

[33] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation and clinical use." *Circulation*, vol. 93, no. 5, pp. 1043–1065, 03 1996. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/8598068

[34] M. Esco and A. Flatt, "Ultra-short-term heart rate variability indexes at rest and post-exercise in athletes: Evaluating the agreement with accepted recommendations," *Journal of sports science & medicine*, vol. 13, 09 2014.

[35] R. Castaldo, L. Pecchia, and P. Melillo, "Acute mental stress detection via ultra-short term hrv analysis," *IFMBE Proceedings*, vol. 51, 01 2015.

[36] L. Pecchia, R. Castaldo, L. Montesinos, and P. Melillo, "Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations," *Healthc Technol Lett*, vol. 5, no. 3, pp. 94–100, Jun 2018.

[37] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term hrv features as surrogates of short term hrv: a case study on mental stress detection in real life," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 12–12, 01 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30654799

[38] J. Shen, J. Barbera, and C. M. Shapiro, "Distinguishing sleepiness and fatigue: focus on definition and measurement," *Sleep Medicine Reviews*, vol. 10, no. 1, pp. 63 – 76, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1087079205000444

[39] M. W. Johns, "A new method for measuring daytime sleepiness: The Epworth Sleepiness Scale." Johns, Murray W.: Sleep Disorders Unit, Epworth Hospital, Melbourne, VIC, Australia, 3121, pp. 540–545, 1991.

[40] "Transport Accident Commission fatigue statistics," https://www.optalert.com/why-optalert/science/#johnsdrowsinessscores, accessed: 2020-09-14.

[41] A. S. et al. (eds.), *STOP, THAT and One Hundred Other Sleep Scales*. Springer Science+Business Media, 2012, ch. 47, pp. 209–210.

[42] J. Horne and S. Baulk, "Awareness of sleepiness when driving," *Psychophysiology*, vol. 41, pp. 161–5, 02 2004.

[43] T. Åkerstedt, B. Peters, A. Anund, and G. Kecklund, "Impaired alertness and performance driving home from the night shift: A driving simulator study," *Journal of sleep research*, vol. 14, pp. 17–20, 04 2005.

[44] P. Philip, P. Sagaspe, N. Moore, J. Taillard, A. Charles, C. Guilleminault, and B. Bioulac, "Fatigue, sleep restriction and driving performance," *Accident Analysis and Prevention*, vol. 37, no. 3, pp. 473–478, 2005.

[45] S. Otmani, T. Pebayle, J. Roge, and A. Muzet, "Effect of driving duration and partial sleep deprivation on subsequent alertness and performance of car drivers," *Physiology and Behavior*, vol. 84, no. 5, 2005.

[46] S. M. Belz, G. S. Robinson, and J. G. Casali, "Temporal separation and self-rating of alertness as indicators of driver fatigue in commercial motor vehicle operators," *Human Factors*, vol. 46, no. 1, pp. 154–169, mar 2004.

[47] J. Axelsson, T. Åkerstedt, G. Kecklund, and A. Lowden, "Tolerance to shift work - how does it relate to sleep and wakefulness?" *International archives of occupational and environmental health*, vol. 77, pp. 121–9, 03 2004.

[48] M. Gillberg, "Subjective alertness and sleep quality in connection with permanent 12-hour day and night shifts," in *Scandinavian Journal of Work, Environment and Health*, vol. 24, no. SUPPL. 3, 1998, pp. 76–80.

[49] M. Härmä, M. Sallinen, R. Ranta, P. Mutanen, and K. Müller, "The effect of an irregular shift system on sleepiness at work in train drivers and railway traffic controllers," *Journal of Sleep Research*, vol. 11, no. 2, 2002.

[50] L. Reyner and J. Horne, "Falling asleep whilst driving: Are drivers aware of prior sleepiness?" *International journal of legal medicine*, vol. 111, pp. 120–3, 02 1998.

[51] K. Kräuchi, C. Cajochen, and A. Wirz-Justice, "Waking up properly: Is there a role of thermoregulation in sleep inertia?" *Journal of Sleep Research*, vol. 13, no. 2, 2004.

[52] M. Gillberg, G. Kecklund, and T. Akerstedt, “Relations between performance and subjective ratings of sleepiness during a night awake,” *Sleep*, vol. 17, no. 3, 1994.

[53] M. Gillberg, G. Kecklund, J. Axelsson, and T. Åkerstedt, “The effects of a short daytime nap after restricted night sleep,” *Sleep*, vol. 19, no. 7, 1996.

[54] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa, “Validation of the karolinska sleepiness scale against performance and eeg variables,” *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 117, pp. 1574–81, 08 2006.

[55] T. Akerstedt and M. Gillberg, “Subjective and objective sleepiness in the active individual,” *Int. J. Neurosci.*, vol. 52, no. 1-2, pp. 29–37, May 1990.

[56] L. Oliveira, J. S. Cardoso, A. Lourenço, and C. Ahlström, “Driver drowsiness detection: a comparison between intrusive and non-intrusive signal acquisition methods,” in *2018 7th European Workshop on Visual Information Processing (EUVIP)*, Nov 2018, pp. 1–6.

[57] C. S. Silveira, J. S. Cardoso, A. L. Lourenço, and C. Ahlström, “Importance of subject-dependent classification and imbalanced distributions in driver sleepiness detection in realistic conditions,” *IET Intelligent Transport Systems*, vol. 13, no. 2, pp. 347–355, 2019.

[58] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, “Stanford Sleepiness Scale (SSS),” in *STOP, THAT and One Hundred Other Sleep Scales*, 2011.

[59] “Stanford sleepiness scale,” https://web.stanford.edu/~dement/sss.html, accessed: 2020-09-17.

[60] T. Åkerstedt, A. Anund, J. Axelsson, and G. Kecklund, “Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function,” *Journal of Sleep Research*, vol. 23, no. 3, 2014.

[61] D. Michie, “”memo” functions and machine learning,” *Nature*, vol. 218, no. 5136, 1968.

[62] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. Verlag New York: Springer, 2009, ch. 2, pp. 201–213.

[63] C. Merow, M. J. Smith, T. C. Edwards Jr, A. Guisan, S. M. McMahon, S. Normand, W. Thuiller, R. O. Wüest, N. E. Zimmermann, and J. Elith, “What do we gain from simplicity versus complexity in species distribution models?” *Ecography*, vol. 37, no. 12, pp. 1267–1281, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.00845

[64] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *International Journal of Advances in Soft Computing and its Applications*, vol. 7, 2015.

[65] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[66] M. Ashtiyani, S. Navaei Lavasani, A. Asgharzadeh Alvar, and M. R. Deevband, "Heart rate variability classification using support vector machine and genetic algorithm," *Journal of biomedical physics & engineering*, vol. 8, no. 4, pp. 423–434, 12 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30568932

[67] M. Babaeian and M. Mozumdar, "Driver drowsiness detection algorithms using electrocardiogram data analysis," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 2019, pp. 0001–0006.

[68] G. Li and W.-Y. Chung, "Detection of driver drowsiness using wavelet analysis of heart rate variability and a support vector machine classifier," *Sensors*, vol. 13, no. 12, pp. 16 494–16 511, 2013. [Online]. Available: https://www.mdpi.com/1424-8220/13/12/16494

[69] T. Xing, Q. Wang, C. Q. Wu, W. Xi, and X. Chen, "Dwatch: A reliable and low-power drowsiness detection system for drivers based on mobile devices," *ACM Trans. Sen. Netw.*, vol. 16, no. 4, Sep. 2020. [Online]. Available: https://doi.org/10.1145/3407899

[70] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: http://www.jstor.org/stable/2699986

[71] D. Graupe, *Principles of Artificial Neural Networks*, 2nd ed. USA: World Scientific Publishing Co., Inc., 2007.

[72] A. J. Ferreira and M. A. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern Recognition Letters*, vol. 33, 2012.

[73] ——, "An unsupervised approach to feature discretization and selection," *Pattern Recognition*, vol. 45, 2012.

[74] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, 2004.

[75] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002.

[76] L. V. D. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, 2008.

[77] P. Jackson, C. Hilditch, A. Holmes, N. Reed, N. Merat, and L. Smith, *Fatigue and road safety: a critical analysis of recent evidence*. Department for Transport, 01 2011.

[78] G. Maycock, *Driver sleepiness as a factor in car and HGV accidents*. Transport Research Laboratory, January 1995.

[79] D. Flatley, L. Reyner, and J. Horne, “Sleep-related crashes on sections of different road types in the uk (1995– 2001),” in *Road Safety Research Report No. 52*. Department for Transport, 2004.

[80] R. Robertson, E. Holmes, and W. Van Laar, *The Facts about Fatigued Driving in Ontario, A Guidebook for Police*. Traffic Injury Research Foundation, 2009.

[81] European Road Safety Observatory, *Fatigue*. European comission, 2018.

[82] National Sleep Foundation, *”Sleep America” Pool 2002*. National Sleep Foundation, 2002.

[83] J. Connor, G. Whitlock, R. Norton, and R. Jackson, “The role of driver sleepiness in car crashes: a systematic review of epidemiological studies,” *Accident; analysis and prevention*, vol. 33, no. 1, pp. 31–41, 01 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11189120

[84] T. a. Dingus, S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. a. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. Knipling, “The 100-car naturalistic driving study phase ii – results of the 100-car field experiment,” *Dot Hs 810 593*, 2006.

[85] J. Herman, B. Kafoa, I. Wainiqolo, E. Robinson, E. McCaig, J. Connor, R. Jackson, and S. Ameratunga, “Driver sleepiness and risk of motor vehicle crash injuries: a population-based case control study in fiji (trip 12),” *Injury*, vol. 45, no. 3, pp. 586–591, 03 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23830198

[86] R. Martinez, “National highway traffic safety administration,” Washington, DC, July 1997.

[87] L. Shamoa-Nir and M. Koslowsky, “Aggression on the road as a function of stress, coping strategies and driver style,” *Psychology*, vol. 01, pp. 35–44, 01 2010.

[88] D. Hennessy and D. Wiesenthal, “The relationship between traffic congestion, driver stress and direct versus indirect coping behaviours,” *Ergonomics*, vol. 40, pp. 348–361, 03 1997.

[89] G. Matthews, L. Dorn, T. Hoyes, D. Davies, I. Glendon, and R. Taylor, “Driver stress and performance on a driving simulator,” *Human factors*, vol. 40, pp. 136–49, 04 1998.

[90] K. Różanowski, O. Truszczyński, K. Filipczak, and M. Madeyski, *The level of driver personality and stress experienced as factors influencing behavior on the road*. WIT Press, 2015.

[91] T. Litman, *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute Victoria, Canada, 2017.

[92] A. Kircher, M. Uddman, and J. Sandin, "Vehicle control and drowsiness," *Swedish National Road and Transport Research Institute*, 2002.

[93] C. D. Wylie, T. Shultz, M. M. Mitler, and R. R. MACKIE, "Commercial motor vehicle driver fatigue and alertness study," in *Technical Summary*. Transport Canada, 1996. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/2212

[94] R. Bittner, K. Hána, L. Poušek, P. Smrka, P. Schreib, and P. Vysoký, "Detecting of fatigue states of a car driver," in *Medical Data Analysis*, R. W. Brause and E. Hanisch, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 260–273.

[95] T. Dingus, L. Hardee, and W. Wierwille, "Detection of drowsy and intoxicated drivers based on highway driving performance measures." in *IEOR Department Report #8402*. Blacksburg, Virginia, USA: Vehicle Simulation Laboratory, Human Factors Group, 1985.

[96] J. Skipper, W. Wierwille, and L. Hardee, "An investigation of low-level stimulus-induced measures of driver drowsiness," in *IEOR Department Report #8402*. Blacksburg, Virginia, USA: Vehicle Simulation Laboratory, Human Factors Group, 1984.

[97] P. Batavia, "Driver-adaptive lane departure warning systems," Ph.D. dissertation, Carnegie Mellon University, The Robotics Institute; Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 9 1999.

[98] W. B. Verwey and D. M. Zaidel, "Predicting drowsiness accidents from personal attributes, eye blinks and ongoing driving behaviour," *Personality and Individual Differences*, vol. 28, no. 1, pp. 123 – 142, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191886999000896

[99] D. Waard, "The measurement of drivers' mental workload," Ph.D. dissertation, University of Groningen, Haren, The Netherlands : University of Groningen, Traffic Research Centre, 1996.

[100] M. Muñoz-Organero and V. Corcoba-Magaña, "Predicting upcoming values of stress while driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1802–1811, July 2017.

[101] Q. Ji and X. Yang, "Real-time eye, gaze, and face pose tracking for monitoring driver vigilance," *Real-Time Imaging*, vol. 8, no. 5, pp. 357 – 377, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077201402902792

[102] W. W. Wierwille, S. S. Wreggit, C. L. Kirn, L. A. Ellsworth, and R. J. Fairbanks, “Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness. final report,” Virginia Polytechnic Institute and State University, Blacksburg; National Highway Traffic Safety Administration Office of Crash Avoidance Research, 1200 New Jersey Avenue, SE Washington, DC United States 20590, Tech. Rep., 12 1994.

[103] L. Bergasa, J. Nuevo, M.-A. Sotelo, R. Barea, and M. Guillén, “Real-time system for monitoring driver vigilance.” *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, pp. 63–77, 01 2006.

[104] P. Smith, M. Shah, and N. da Vitoria Lobo, “Determining driver visual attention with one camera,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 4, pp. 205–218, Dec 2003.

[105] G. Masala and E. Grosso, “Real time detection of driver attention: Emerging solutions based on robust iconic classifiers and dictionary of poses,” *Transportation Research Part C: Emerging Technologies*, vol. 49, pp. 32 – 42, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X14002976

[106] L. Zhang, F. Liu, and J. Tang, “Real-time system for driver fatigue detection by rgb-d camera,” *ACM Transactions on Intelligent Systems and Technology*, vol. 6, pp. 1–17, 03 2015.

[107] R. K. Malhotra and A. Y. Avidan, “Chapter 3 - sleep stages and scoring technique,” in *Atlas of Sleep Medicine (Second Edition)*, 2nd ed., S. Chokroverty and R. J. Thomas, Eds. St. Louis: W.B. Saunders, 2014, pp. 77 – 99. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9781455712670000035

[108] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu, “Recognizing the degree of human attention using EEG signals from mobile sensors,” *Sensors (Basel, Switzerland)*, vol. 13, no. 8, pp. 10 273–10 286, aug 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23939584https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3812603/

[109] Chin-Teng Lin, Ruei-Cheng Wu, Sheng-Fu Liang, Wen-Hung Chao, Yu-Jie Chen, and Tzyy-Ping Jung, “Eeg-based drowsiness estimation for safety driving using independent component analysis,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 12, pp. 2726–2738, Dec 2005.

[110] Ruey S. Huang, Chung J. Kuo, Ling-Ling Tsai, and O. T. C. Chen, “Eeg pattern recognition-arousal states detection and classification,” in *Proceedings of International Conference on Neural Networks (ICNN’96)*, vol. 2, June 1996, pp. 641–646 vol.2.

[111] A. Vuckovic, V. Radivojevic, A. C. Chen, and D. Popovic, “Automatic recognition of alertness and drowsiness from eeg by an artificial neural network,” *Medical Engineering & Physics*, vol. 24, no. 5, pp. 349 – 360, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350453302000309

[112] B. Wilson and T. Bracewell, “Alertness monitor using neural networks for eeg analysis,” in *Proceedings of the 2000 IEEE Signal Processing Society Workshop*, vol. 2, 02 2000, pp. 814 – 820 vol.2.

[113] H. D. Critchley, R. Elliott, C. J. Mathias, and R. J. Dolan, “Neural activity relating to generation and representation of galvanic skin conductance responses: a functional magnetic resonance imaging study,” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 20, no. 8, pp. 3033–3040, apr 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10751455https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6772223/

[114] J. Healey and R. Picard, “Smartcar: detecting driver stress,” in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 4, 2000, pp. 218–221 vol.4.

[115] J. A. Healey and R. W. Picard, “Detecting stress during real-world driving tasks using physiological sensors,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[116] A. Chowdhury, R. Shankaran, M. Kavakli, and M. M. Haque, “Sensor applications and physiological features in drivers’ drowsiness detection: A review,” *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3055–3067, 2018.

[117] C. S. A. R. da Silva Silveira, “Driver’s Fatigue State Monitoring using Physiological Signals,” Master’s thesis, Universidade do Porto - Faculdade de Engenharia, Porto, Portugal, 2017.

[118] G. Rigas, Y. Goletsis, P. Bougia, and D. I. Fotiadis, “Towards Driver’s State Recognition on Real Driving Conditions,” *International Journal of Vehicular Technology*, vol. 2011, p. 617210, 2011. [Online]. Available: https://doi.org/10.1155/2011/617210

[119] M. J. C. N. Fernandes, “Driver drowsiness detection using non-intrusive eletrocardiogram and steering wheel angle signals,” Master’s thesis, Universidade do Porto - Faculdade de Engenharia, Porto, Portugal, 2019.

[120] T. Gruden, K. Stojmenova, J. Sodnik, and G. Jakus, “Assessing drivers’ physiological responses using consumer grade devices,” *Applied Sciences*, vol. 9, p. 5353, 12 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/24/5353

[121] A. Lourenço, A. P. Alves, C. Carreiras, R. P. Duarte, and A. Fred, “Cardiowheel: Ecg biometrics on the steering wheel,” in *Machine Learning and Knowledge Discovery in Databases*, A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, and M. Spiliopoulou, Eds. Cham: Springer International Publishing, 2015, pp. 267–270.

[122] A. Lourenço, H. Silva, C. Carreiras, , and Fred, “Outlier detection in non-intrusive ecg biometric system,” in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7950 LNCS, 2013.

[123] S. H.-W. Chuah, P. A. Rauschnabel, N. Krey, B. Nguyen, T. Ramayah, and S. Lade, “Wearable technologies: The role of usefulness and visibility in smartwatch adoption,” *Computers in Human Behavior*, vol. 65, pp. 276–284, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563216305374

[124] J. Pan and W. J. Tompkins, “A real-time qrs detection algorithm,” *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[125] H. Tanaka, K. D. Monahan, and D. R. Seals, “Age-predicted maximal heart rate revisited,” *J Am Coll Cardiol*, vol. 37, no. 1, pp. 153–156, Jan 2001.

[126] J. W. Mason, D. J. Ramseth, D. O. Chanter, T. E. Moon, D. B. Goodman, and B. Mendzelevski, “Electrocardiographic reference ranges derived from 79,743 ambulatory subjects,” *Journal of Electrocardiology*, vol. 40, no. 3, pp. 228 – 234.e8, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022073606003530

[127] C. Fors, C. Ahlstrom, P. Sorner, J. Kovaceva, E. Hasselberg, M. Krantz, J. F. Gronvall, K. Kircher, and A. Anund, “Camera-based sleepiness detection: final report of the project sleepeye.” *Vip Publication*, 2011.

[128] R. E. Kleiger, J. Miller, J. Bigger, and A. J. Moss, “Decreased heart rate variability and its association with increased mortality after acute myocardial infarction,” *The American Journal of Cardiology*, vol. 59, no. 4, pp. 256 – 262, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0002914987907958

[129] A. Persson, H. Jonasson, I. Fredriksson, U. Wiklund, and C. Ahlström, “Heart rate variability for driver sleepiness classification in real road driving conditions*,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 6537–6540.

[130] M. Nardelli, A. Greco, J. Bolea, G. Valenza, E. Scilingo, and R. Bailón, “Reliability of lagged poincaré plot parameters in ultrashort heart rate variability series: Application on affective sounds,” *IEEE Journal of Biomedical and Health Informatics*, vol. PP, pp. 1–1, 04 2017.

[131] P. Costa, “Simulador de tráfego para monitorização e assistência à condução,” Bachelor’s Thesis, Instituto Superior de Engenharia de Lisboa, Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores; Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal, 9 2020.

[132] R. M. S. Zambujal, “Fadiga ocupacional e processos de regulação emocional: Um estudo exploratório com tripulantes de cabine,” Master’s thesis, Instituto Universitário de ciências Psicológicas, Sociais e da Vida, ISPA, R. Jardim do Tabaco 34, 1100-304 Lisboa, Portugal, 2013.

[133] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947. [Online]. Available: https://doi.org/10.1007/BF02295996

[134] M. P. Fay, “Exact mcnemar’s test and matching confidence intervals,” *R*, 2015.