



PEDRO DE BASTOS VEIGA

Bachelor of Science in Biomedical Engineering

MACHINE LEARNING ALGORITHM DEVELOPMENT OF SPO2 SENSOR FOR IMPROVED ROBUSTNESS IN WEARABLES

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
December, 2022

MACHINE LEARNING ALGORITHM DEVELOPMENT OF SPO2 SENSOR FOR IMPROVED ROBUSTNESS IN WEARABLES

PEDRO DE BASTOS VEIGA

Bachelor of Science in Biomedical Engineering

Adviser: Prof. Dr. Hugo Filipe Silveira Gamboa
Associate Professor, NOVA University Lisbon

Examination Committee

Chair: Prof. Dr. Célia Maria Reis Henriques
Associate Professor, NOVA University Lisbon

Rapporteur: Prof. Dr. Valentina Borissovna Vassilenko
Assistant Professor, NOVA University Lisbon

Adviser: Prof. Dr. Hugo Filipe Silveira Gamboa
Associate Professor, NOVA University Lisbon

Machine learning algorithm development of SPO2 sensor for improved robustness in wearables

Copyright © Pedro de Bastos Veiga, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

Acknowledgements

First of all, I would like to thank my adviser, Professor Hugo Gamboa, for giving me the opportunity to explore such an interesting topic that helped me to develop and improve soft and hard skill in many subjects and allowed to become a better Biomedical Engineer.

To Rui Varandas, who was always available during all these months and supported me and taught me so many things. Without him, I would not have developed so much, in both the work presented and myself. Thank you for your professionalism and your friendship, I could not have asked for a better mentor.

I would also like to thank PLUX Wireless Biosignals for believing in me and giving me the opportunity to develop something that could be implemented in a future device.

To FCT NOVA, my home for the last 5 years, thank you for making me grow into the person that I am today. To my friends in Micro and Nano, thank you for all your love and support, you will always be in my heart. To my friends in Biomedical Engineering, thank you for the last 4 years of struggle, support and, most importantly, good times.

To Mané, my godson, roommate and, more than anything, friend, thank you for all your support. Thank you for always being there when I needed it and helping me with my problems. Thanks for your patience during all these years.

To my family, thank you for dealing with me for 23 years, I know it is not easy. To my parents, thank you for all the education you gave me, you made me a better person. To Inês, my little sister, thank you for all your love and affection.

Finally, but definitively not least, a big thank you for my beloved Fabiana. Thank you for, even 380 km away, brightening my days, making me smile and surprising me everyday. You are the most caring person I have ever met.

*“ How can we ever be justified in generalizing
from what we’ve seen to what we haven’t? ”*
(David Hume)

Abstract

Wearable devices application in the digital measurement of health has gained attention by researchers. These devices allow for data acquisition during real-life activities, resulting in higher data availability. They often include photoplethysmography (PPG) sensors, the sensor behind pulse oximetry. Pulse oximetry is a non-invasive method for continuous oxygen saturation (SpO_2) measurements, a standard monitor for anesthesia procedures, an essential tool for managing patients undergoing pulmonary rehabilitation and an effective method for assessing sleep-disordered breathing. However, the current market focuses on heart rate measurements and lacks the robustness of clinical applications for SpO_2 assessment. In addition, the most common obstacle in PPG measurements is the signal quality, especially in the form of motion artifacts. Thus, this work aims at increasing the clinical robustness in this devices by evaluating its quality and then extracting relevant metrics.

Firstly, a data acquisition protocol was developed, focused on acquiring data during daily activities. This resulted in a dataset with different signal qualities, which was manually annotated to be used as the base for the Machine Learning models. A second protocol was also developed especially designed for the extraction of the SpO_2 measurement.

Several Machine Learning models were developed to evaluate the signal in three distinct qualities (corrupted, suboptimal, optimal) in real time. A Random Forest classifier achieved accuracies of 79% and 80% for the binary models capable of differentiating between usable and unusable signals, and accuracies of 74% and 80% when distinguishing between optimal and suboptimal signals, for the two utilized channels. The multi-class models achieved accuracies of 66% and 65% for the two utilized channels.

Three clinically relevant metrics were also extracted from the PPG signal: heart rate, respiratory rate and SpO_2 . The heart rate and respiratory rate algorithms resulted in performances similar to the ones found in the literature and in other devices currently on the market. However, while promising, more data is needed to reach statistical significance for the SpO_2 measurement.

Keywords: Photoplethysmography, Pulse Oximetry, Machine Learning, Signal Quality, Heart Rate, Respiratory Rate, SpO_2

Resumo

A monitorização do estado de saúde de pacientes em ambulatório utilizando dispositivos *wearables* tem vindo a ser cada vez mais investigada. Estes dispositivos permitem uma aquisição de dados durante o dia a dia, resultando num maior conjunto de dados. Frequentemente, estes dispositivos incluem fotopletismógrafos (PPG), o sensor por detrás da oximetria de pulso. A oximetria de pulso é um método não invasivo para a medição da saturação de oxigénio no sangue (SpO_2) de forma contínua. É um equipamento padrão para procedimentos com anestesia, uma ferramenta essencial para monitorizar pacientes em reabilitação pulmonar e um método eficaz para avaliar respiração desordenada do sono. Ainda assim, o mercado atual foca-se principalmente em medições da frequência cardíaca e carece robustez para aplicações clínicas da medição de SpO_2 . Para além disso, o obstáculo mais comum em medições com PPG é a qualidade do sinal. Consequentemente, este trabalho procura melhorar a robustez clínica destes dispositivos analisando a qualidade do sinal e, posteriormente, extrair métricas relevantes.

Primeiramente, foi desenvolvido um protocolo para aquisição de dados de atividades do dia a dia. Assim, foram adquiridos dados com diferentes qualidades, que foram avaliados manualmente de forma a servir de base para os vários modelos de *Machine Learning*. Também foi desenvolvido um segundo protocolo para a extração do valor de SpO_2 .

Diferentes modelos de *Machine Learning* foram desenvolvidos para avaliar em tempo real a qualidade do sinal em três qualidades (corrompido, subótimo, ótimo). Um classificador baseado em *Random Forest* atingiu exatidões de 79% e 80% em classificadores binários capazes de distinguir entre sinais úteis e inúteis, e exatidões de 74% e 80% a diferenciar entre um sinal subótimo e ótimo, para os dois canais usados. Os classificadores multi-classe atingiram exatidões de 66% e 65% para os dois canais usados.

Três medidas clinicamente relevantes foram também extraídas do sinal de PPG: frequências cardíaca e respiratória, cujos algoritmos atingiram resultados semelhantes aos encontrados na literatura e em aparelhos no mercado, e SpO_2 que, ainda que promissores, mais dados seriam necessários para os resultados serem estatisticamente significativos.

Palavras-chave: Fotopletismografia, Oximetria de Pulso, *Machine Learning*, Qualidade do Sinal, Frequência Cardíaca, Frequência Respiratória, SpO_2

Contents

List of Figures	x
List of Tables	xii
Acronyms	xv
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	2
1.3 Structure	3
2 Theoretical Concepts	4
2.1 Photoplethysmography	5
2.1.1 Light Source	6
2.1.2 Sensor Location	7
2.1.3 Limitations	8
2.2 SpO ₂ Calculation	9
2.3 Machine Learning	10
2.3.1 Decision Trees	11
2.3.2 Bagging	12
2.3.3 Train-Test Sets	13
2.3.4 Performance Evaluation	14
2.3.5 Under and Overfitting	18
2.3.6 Balancing Data	18
3 State of the Art	19
3.1 Photoplethysmography	19
3.1.1 LED Count and Wavelengths	19
3.1.2 Sensor Location	20
3.1.3 Quality Assessement	21

3.1.4	Artifacts Removal	21
3.2	Clinical Applications	23
3.2.1	Blood Pressure	23
3.2.2	SpO ₂	23
3.2.3	Carboxyhemoglobin Saturation	24
3.2.4	Respiratory Rate	24
3.2.5	Heart Rate	25
4	Materials and Methods	27
4.1	Data Acquisition	27
4.1.1	Acquisition Protocol for Quality Assessment	29
4.1.2	Acquisition Protocol for SpO ₂ Extraction	29
4.2	Data Processing	30
4.2.1	Digital Filter	30
4.2.2	Task and Block Division	32
4.2.3	Quality Assessment	32
4.2.4	Normalization and Standardization	35
4.3	Feature Extraction	35
4.3.1	Feature Selection	37
4.4	Machine Learning Models	38
4.5	Physiological Metrics	38
4.5.1	Heart Rate Extraction	38
4.5.2	Respiratory Rate Extraction	39
4.5.3	SpO ₂ Extraction	40
4.5.4	Metrics Evaluation	42
5	Results	44
5.1	Data Acquisition	44
5.1.1	Acquisition Protocol for Quality Assessment	44
5.1.2	Acquisition Protocol for SpO ₂ Extraction	44
5.2	Quality Assessment	45
5.2.1	Regression Models	46
5.2.2	Classification Models	48
5.2.3	Feature Importance	57
5.3	Physiological Metrics	59
5.3.1	Heart Rate Extraction	59
5.3.2	Respiratory Rate Extraction	61
5.3.3	SpO ₂ Extraction	63
6	Discussion	66
6.1	Quality Assessment	66
6.1.1	Manual Classification Agreement	66

6.1.2	Regression Models	67
6.1.3	Classification Models	67
6.1.4	Feature Importance	70
6.2	Physiological Metrics	71
6.2.1	Heart Rate Extraction	71
6.2.2	Respiratory Rate Extraction	72
6.2.3	SpO ₂ Extraction	73
7	Conclusion	74
7.1	Future Work	75
	Bibliography	77
	Appendices	
A	Informed Consent for Data Acquisition	88

List of Figures

1.1	Results for the "(pulse oximet*) AND (wearable)" query in PubMed.	2
2.1	Sources of light attenuation in tissue and blood and respective PPG waveform.	5
2.2	LED and photodetector (PD) placement for both modes of PPG.	6
2.3	Extinction coefficient curves for HbO ₂ and Hb.	7
2.4	LED cycle diagram showing the on/off switches and respective sampling.	7
2.5	Example of a R-SpO ₂ calibration curve.	10
2.6	Decision tree example with the correct terminology, using the Iris dataset.	12
2.7	Error curve for bagging 1 to 200 decision trees, compared with a single decision tree.	13
2.8	3-fold random subsampling (a) and 3-fold cross validation (b) example.	15
2.9	Subject-wise cross-validation example.	15
2.10	Underfitting example (Linear model) and overfitting example (10 th order model) of a known function.	18
3.1	Sample PPG segments from three different classes.	22
3.2	Block diagram of an adaptive noise canceling algorithm.	22
3.3	Respiratory-induced variations on the PPG signal.	25
4.1	biosignalsplux 8-port hub (a), SpO ₂ sensor (b), inductive based respiration sensor (c) and ECG sensor (d).	28
4.2	Sensors placement.	28
4.3	Experimental protocol for quality assessment.	29
4.4	Experimental protocol for SpO ₂ extraction.	30
4.5	Contec CMS50D+ Pulse Oximeter.	30
4.6	Data processing workflow.	31
4.7	Raw and filtered red channel signal example.	31
4.8	Schematic of a sliding window with overlap.	32
4.9	Data quality examples.	34
4.10	Example of the quality assessment of a 3 s block with a 1.4 average.	35

4.11	Histograms with positive, zero and negative skewness.	36
4.12	Histograms with positive, zero and negative kurtosis.	37
4.13	RIP signal with the zero-crossings highlighted.	40
4.14	RIAV, RIIV and RIFV processing for RR estimation.	40
4.15	PPG signal with the respiratory-induced variations, and respective FFTs. .	41
4.16	Bland-Altman plot for HR agreement between ECG and PPG.	43
5.1	SpO ₂ evolution during a breathing exercise.	45
5.2	RMSE results for the red and IR channels in function of the number of trees generated and the maximum depth.	47
5.3	Real and predicted red quality for the best red channel regressor.	48
5.4	Accuracy results for the red and IR channels in function of the number of trees generated and the maximum depth. Models developed using all features. .	50
5.5	Confusion matrices for the best classification models for the red and IR channels, using the two different quality approaches and features from both channels.	51
5.6	Accuracy results for the red and IR channels in function of the number of trees generated and the maximum depth. Models developed using only features from the individual channel.	52
5.7	Confusion matrices for the best classification models for the red and IR channels, using the two different quality approaches and features from individual channels.	53
5.8	Confusion matrices for the best classification models for the red and IR channels for the double classifiers.	55
5.9	Confusion matrices for the best vote classification models for the red and IR channels, using the two different quality approaches and features from both channels.	57
5.10	Correlation between all the features and the quality from the red and IR channels, for both quality interpretations.	58
5.11	Accuracy for models trained with different amounts of features.	59
5.12	Bland-Altman plot for HR estimation, using the ' <i>Simulation</i> ' approach and the 0.5 Hz-5 Hz band-pass filter.	61
5.13	Bland-Altman plot for RR estimation, using the ' <i>Simulation</i> ' approach and the original filter.	62
5.14	RMSE evolution with the increase of the tuning parameter λ	63
5.15	SpO ₂ calibration curve for the proposed expression.	64
5.16	Bland-Altman plot for SpO ₂ estimation.	65

List of Tables

2.1	Models tested for classification and regression of the data quality.	11
2.2	Confusion matrix for binary classification.	16
3.1	LEDs used and their mode of operation in SpO ₂ devices, where IR represents an infrared LED.	19
4.1	Personal information collected in the survey.	29
4.2	Signal used for each extracted feature.	36
5.1	Results of the manual quality evaluation indicating the prevalence of the different interpretations.	45
5.2	Results of the manual quality evaluation, by subject.	46
5.3	Performance results for the different regression models.	46
5.4	Performance results for the best combinations for the regression models. .	48
5.5	Performance results for the different classification models.	49
5.6	Performance results for the best combinations for the classification models, using features from both channels.	49
5.7	Performance results for the best combinations for the classification models, using features from individual channels.	53
5.8	Performance results for the best combinations for the classification models differentiating between class '0' and '1 $\cup$ 2', using features from both channels.	54
5.9	Performance results for the best combinations for the classification models differentiating between class '1' and '2', using features from both channels.	54
5.10	Performance results for the best combinations of the vote classification models, using features from both channels.	56
5.11	Results for the three approaches for HR estimation with at least one channel with quality '1' or better.	60
5.12	Results for the three approaches for HR estimation with at least one channel with quality '2'.	60
5.13	Performance results for the ' <i>Simulation</i> ' approach for HR estimation. . . .	60

5.14 Performance results for the three approaches for the RR estimation.	62
5.15 Performance results for the ' <i>Simulation</i> ' approach for the RR estimation. . .	62
5.16 Results for the best SpO ₂ extraction models.	64

Acronyms

AC	alternated current
ANC	adaptive noise cancellation
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BP	blood pressure
bpm	beats/min
brpm	breaths/min
COHb	carboxyhemoglobin
DC	direct current
ECG	electrocardiogram
FFT	fast Fourier transform
GPS	Global Positioning System
Hb	deoxyhemoglobin
HbO₂	oxyhemoglobin
HR	heart rate
ICA	independent component analysis
IR	infrared
LED	light-emitting diode
MA	motion artifact
MAE	mean absolute error

MAPE	mean absolute percent error
MetHb	methemoglobin
ML	machine learning
PD	photodetector
PPG	photoplethysmography
PTP	peak-to-peak
RF	Random Forest
RIAV	respiratory-induced amplitude variation
RIFV	respiratory-induced frequency variation
RIIV	respiratory-induced intensity variation
RIP	inductive respiration
RIV	respiratory-induced variations
RMSE	root mean square error
RR	respiratory rate
SaO₂	arterial blood oxygen saturation
SHb	sulfhemoglobin
SpO₂	peripheral oxygen saturation
SQI	signal quality index
WHM	Wim Hof Method

Introduction

1.1 Motivation

Wearable technology has been one of the fitness top trends in the last years, as reported by the "Worldwide Survey of Fitness Trends for 2022" [2]. With the advancement of technology in the last years, companies have developed and promoted wearable products with a mass public attention. In 2021, it was estimated that there will be over 1 billion wearables worldwide by 2022 [3] with this industry being evaluated in US\$100 billion [2]. In addition, the COVID-19 pandemic increased the demand for remote monitoring [4], which is reported to improve quality of life, as well as reducing outpatient care costs by 50% [5].

These devices are gaining attention by researchers for their potential contribution to a digital measurement of health, as they allow for acquisition of patient information during real-life activities, resulting in more representative datasets [5–7], even though it could be at the cost of its quality [7–9].

They often include a large group of sensors such as accelerometers, Global Positioning System (GPS), gyroscope and photoplethysmography (PPG). It is reported that all modern smartwatches have an accelerometer and 71% of consumer wearables are equipped with a PPG sensor [10], the sensor behind pulse oximetry.

Pulse oximetry is a non-invasive method for continuous oxygen saturation measurements [11]. It is a standard monitor for all anesthesia procedures in most developed countries and also used in emergency departments and ambulances to assess blood oxygenation [12, 13]. Pulse oximetry is also an essential tool for managing patients undergoing pulmonary rehabilitation, which is a safe and effective technique in a remote setting [14], and for assessing sleep-disordered breathing [15]. Due to its simplicity, it is a relatively low cost device and an easy-to-use tool, therefore, a natural application is in wearable devices. However, the current wearable market focuses on heart rate measurement and lacks the robustness of clinical applications for SpO₂ measurements and, as a result, most of them do not have medical certification [12]. It is also important to note that many of the pulse oximeters currently being sold for home use are placed on

the patient's fingers, with a clip-on design [15, 16]. While this placement is optimal for a wide variety of applications, it has several limitations for extended uses, in particular, this solution can be too obtrusive [15, 17]. A potential solution for pulse oximeters used continuously is a wrist-worn device.

However, this type of devices comes with the cost of signal quality. The most common obstacle in PPG measurements are motion artifacts (MAs) [18], hence, there is an increasing need for a solution to efficiently detect and remove, or reduce, signal noise and distortion [16].

It is also possible to notice a growth in research in this area. By doing a query with the key concepts, "(pulse oximet*) AND (wearable)", it is possible to see a positive trend in the last years, as shown in Figure 1.1.

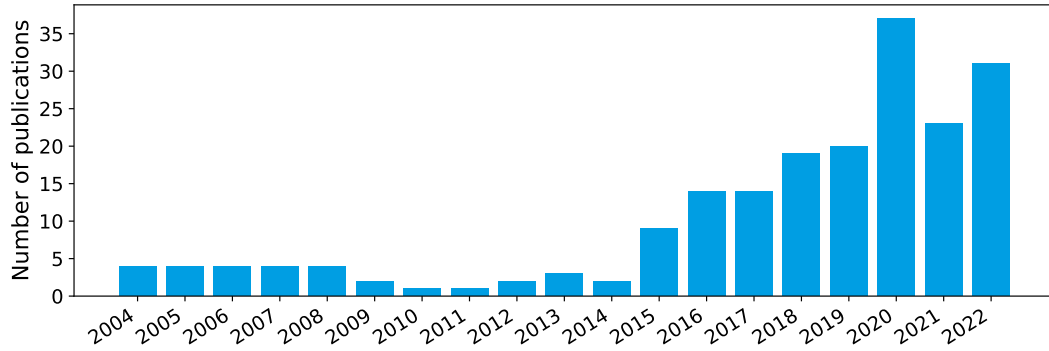


Figure 1.1: Results for the "(pulse oximet*) AND (wearable)" query in PubMed [19].

1.2 Objectives

Based on the statements made above, this thesis aims to develop a solution that allows the improvement of pulse oximetry in wearable devices. This could be divided into three separate objectives:

- Develop a solution that can detect the signal quality in real-time;
- Extract several metrics, such as, heart rate, respiration rate and SpO_2 ;
- Ensure a solution with minimal use of sensors and computational power, for a possible implementation in a wearable device.

Most of the current research uses binary classification and more than one sensor for assessing the PPG signal quality. In this work, a new approach for this problem is used as a single sensor will be used to evaluate the signal into three different qualities. Afterwards, this assessment allows for a differentiated processing whether the signal has sufficient quality for extracting all, some or none biological metrics.

1.3 Structure

This dissertation is divided into 7 chapters. Chapter 2 explains the main theoretical concepts used in this thesis, in particular the PPG sensor and signal, the SpO_2 calculation and base machine learning (ML) notions. A review of the literature concerning PPG sensors, signal quality assessment, artifacts removal and clinical applications are presented in Chapter 3. Chapter 4 includes the material and methods used, including: data acquisition protocols, data processing methods, ML models used and the biological metrics extracted from the signal. The results achieved are presented in Chapter 5, while Chapter 6 discusses them. Finally, Chapter 7 presents the main findings and limitations of this dissertation, along with future work.

Theoretical Concepts

The principal mechanism for oxygen transportation from the lungs to the cells of the body is via hemoglobin molecules in red blood cells. Hemoglobin can be divided into two forms: normal hemoglobin, that is able to bind with O_2 molecules; and dyshemoglobin, which is not. Normal hemoglobin can be sub-divided into deoxyhemoglobin (Hb) and oxyhemoglobin (HbO_2), and dyshemoglobin can be divided into carboxyhemoglobin (COHb), methemoglobin (MetHb), and sulfhemoglobin (SHb) [11].

Arterial blood oxygen saturation (SaO_2) can be determined by the ratio of HbO_2 concentration to total hemoglobin concentration. Usually, only oxygen-carrying hemoglobin are taken into account for the calculation of blood oxygen content, as 97% - 98% of the total oxygen content is carried by it, this approximation is shown in expression 2.1.

$$SaO_2 = \frac{[HbO_2]}{[HbO_2] + [Hb] + [COHb] + [MetHb] + [SHb]} = \frac{[HbO_2]}{[TotalHb]} \approx \frac{[HbO_2]}{[HbO_2] + [Hb]} \quad (2.1)$$

This value, for healthy adults at sea level, will range between 96% and 98% [20]. SaO_2 can be directly obtained by drawing blood from an artery and subjecting it to blood gas analysis. This is a time-consuming and invasive process, that can only be extracted intermittently [21]. In order to measure oxygen saturation noninvasively, Dr. Takuo Aoyagi invented pulse oximetry in 1974 [22]. This method does not measure the SaO_2 directly, but the peripheral oxygen saturation (SpO_2), although they are connected.

For the past 30 years, pulse oximetry has been widely used as a continuous noninvasive monitoring technique of the oxygen saturation of hemoglobin [20], which is directly correlated with oxygen availability. This is an important parameter for evaluating hypoxemia, low blood oxygen level, and the increased use of pulse oximeters can provide an early warning of hypoxemia, therefore it is used during anesthesia, post-operatively, intensive care and sleep laboratories [11, 15]. It can be also used to diagnosis sleep apnea, a condition where the subject has moments where he stops breathing during sleep.

2.1 Photoplethysmography

Pulse oximetry is based on photoplethysmography (PPG). A plethysmograph is a device consisting of a light source and a detector, in which the light source is usually one or more light-emitting diodes (LEDs) of different wavelengths and intensities. This technique is able to measure volume changes in blood vessels and estimate vital signs, such as heart rate (HR) and its variability, respiratory rate (RR), blood pressure (BP), and SpO_2 [23].

The principle behind PPG is the Beer-Lambert law [3], shown in equation 2.2. The Beer-Lambert law describes the transmitted light intensity (I) through a material as a function of the incident light intensity (I_0), the extinction coefficient (ϵ), the concentration of the substance ($[C]$) and the light path length (d). Therefore light traveling through biological tissue is assumed to have the same I_0 and d , making the only variable the $\epsilon[C]$ component, which is specific for different substances. Hence incident light is absorbed differently by different substances, such as skin pigments, muscle, bone, and arterial and venous blood.

$$I = I_0 e^{-\epsilon[C]d} \quad (2.2)$$

Most of these substances remain unchanged, that is, their light attenuation remains constant, therefore, arterial blood in arteries and arterioles is the main factor of changes in the detected light intensity. Venous pulsations can also contribute to changes in light absorption, however, there is the assumption that this parameter does not change considerably [11, 24]. This results in a signal with a direct current (DC) component and an alternated current (AC) component, as seen in Figure 2.1.

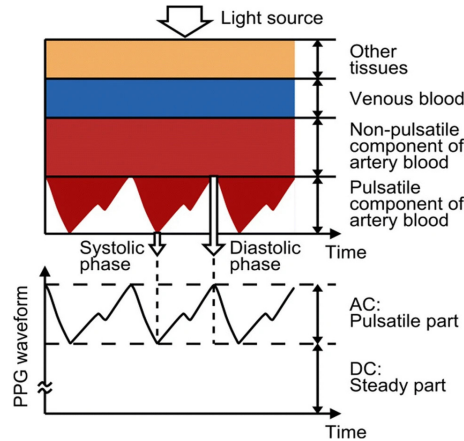


Figure 2.1: Sources of light attenuation in tissue and blood and respective PPG waveform. Reprint from [24].

A PPG device has 2 modes of operation, transmission and reflectance, as shown in Figure 2.2. In transmission mode, the LED and the photodetector (PD) are placed in opposite sides of the measurement site, while in reflectance mode the LED and the photodetector

are placed in the same side, hence the PD detects backscattered light from the body. Currently, PPG sensor in a transmissive mode placed in the finger is the most common use due to the high signal amplitude, high accuracy and stability, but this placement is not convenient for long-term use as most daily activities require the use of the fingers [11, 25]. On the other hand, the main advantage for reflectance mode pulse oximeters is their ability to be placed at nearly any place on the body [26], e.g, wrist and forehead. However the accuracy for this type of sensors can highly vary based on, for instance, ambient light interferences, contact pressure and skin tone [26].



Figure 2.2: LED and photodetector (PD) placement for both modes of PPG, [24].

In wearable devices, a PPG sensor in reflectance mode may be more appropriate, as it can be placed in most parts of the body, but it is important to be cautious regarding the signal quality and its artifacts.

2.1.1 Light Source

Light interaction with the human body can be very complex and involves phenomena such as scattering, absorption and reflection and, even though the widespread use of the PPG, the origin of its signal is still subject of debate [3]. Nevertheless, the photoplethysmography utilizes the light absorption variation in the blood vessels during the cardiac cycle. During the systole there is an increase in blood volume, resulting in an increased absorbance, when compared to the diastolic phase [3]. The two main absorbers in the blood, HbO_2 and Hb , have different absorption coefficients, relative to the wavelength of the incident light, as illustrated in Figure 2.3. This is the concept behind pulse oximetry. The pulse oximeter requires light with at least two distinct wavelengths in such a way that the extinction coefficient of the two oxygen-carrying hemoglobins is different [3].

Currently, most of the PPG based pulse oximeters use a combination of red and infrared (IR) LEDs, with some variations on the exact wavelength, usually using red light in the spectrum of 650-750 nm and IR light in the spectrum of 850-1000 nm [11, 23, 24, 28]. Recent studies are investigating the use of different wavelengths combinations such as green-orange [23], two IR LEDs [28], green-red-IR [15] and a single red LED [29]. Their advantages and disadvantages are mentioned in Section 3.1.1.

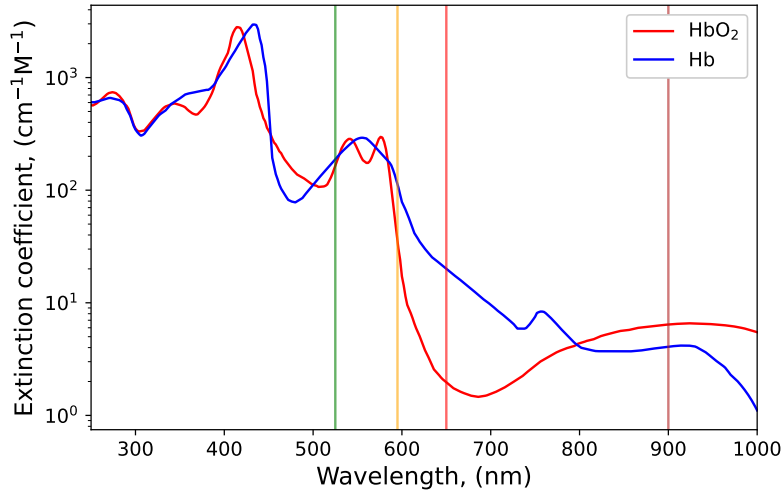


Figure 2.3: Extinction coefficient curves for HbO_2 and Hb , from [27]. Vertical lines in the most frequently used wavelengths in the literature: Green (525 nm); Orange (595 nm); Red (650 nm); Infrared (900 nm).

2.1.1.1 LED cycle

Multiple wavelengths cannot be activated at the same time because they will interfere with each other. Therefore there is the need to turn on and off the LEDs alternately at very high frequencies, typically between a few hundreds to a few thousand hertz [3, 30]. Afterwards, it is possible to sample the two LEDs at different times obtaining two independent signals. There is also the possibility to collect the ambient light, i.e., when both LEDs are turned off. This process can be seen in Figure 2.4.

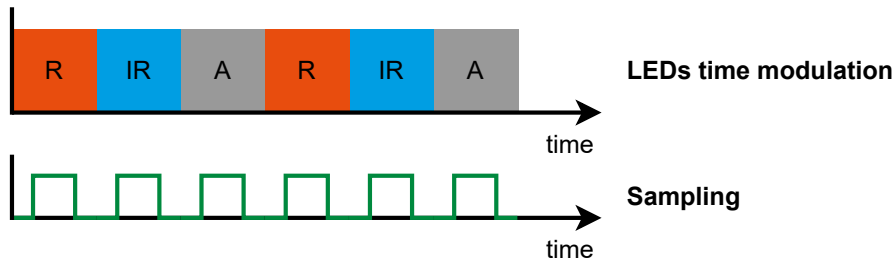


Figure 2.4: LED cycle diagram showing the on/off switches and respective sampling. On top, the on/off times for red (R), infrared (IR) and ambient light (A, with both LEDs switched off). Adapted from [3].

2.1.2 Sensor Location

The most common placement for transmission pulse oximeters is the finger, but there is also the possibility for placement in the earlobe [31]. These sensors need to be placed in well perfused tissues and are prone to motion artifacts [31]. By contrast, the main advantage for reflectance mode pulse oximeters is their ability to be placed at nearly any

place on the body [26], e.g., wrist and forehead. However, the accuracy for this type of sensors can vary greatly based on, for instance, ambient light interferences, contact pressure and skin tone [26]. This subject is discussed in more depth in Section 3.1.2.

2.1.3 Limitations

Even though a pulse oximeter is an essential monitoring and diagnostic device in a hospital environment, its accuracy can be highly affected by a number of elements. These factors can be subject-specific, such as skin pigmentation, nail polish, intravenous dyes, or the high presence of dyshemoglobins (which were excluded from the calculations of SpO_2 , shown in Equation 2.1) [32]. Low perfusion, bright external lights and electromagnetic interference can also cause low signal-to-noise ratio (SNR) [32].

2.1.3.1 Motion Artifacts

Apart from all these factors, the most common and difficult limitation to overcome is MAs [18, 32]. According to Seok et al. [33], a complete PPG signal can be represented as the the sum of several components:

$$Signal = N_{amb} + N_{mech} + N_{elec} + N_{vas} + S_{vas} \approx N_{ma} + S_{vas} \quad (2.3)$$

Where N_{amb} , N_{mech} , N_{elec} , N_{vas} and S_{vas} represents the effects of environmental light, movement, electrical noise, vascular dynamics and the signal of interest created by blood pulsation, respectively. The noise created by N_{amb} and N_{elec} can be excluded via proper circuit design. N_{mech} and N_{vas} are the noise related to motion and the sum of these can be considered as MAs (N_{ma}).

The mechanisms behind these are not fully understood, and the independence of PPG and MA is still a point of controversy [32–34]. Some studies claim that motion might affect arterial blood volume variation [33, 35]. Others argue that motion can cause venous blood movement, in addition to the expected arterial blood [32]. This leads the pulsatile component to be composed of more than just arterial blood. This effect is more notable when combined with low perfusion, where the venous blood pulsations component has a greater contribution [32].

These artifacts, caused by voluntary or involuntary movement, can hide the real signal with noise, making its interpretation impossible. In a clinical setting, this can increase the number of false alarms, diminishing their importance and increasing caregiver workload, stress, and patient care [32]. In a 2015 report on health technology hazards by the ECRI Institute, "Alarm Hazards" was the top hazard, followed by "Data Integrity" [36]. Both of these dangers are increased when MAs are present. Several studies reported that only 23–29% of the alarms were true [32, 37] and, in a infant study, motion artifacts were present in 91% of the awakened time [38]. Charlton et al. [39] investigated the acquisition of PPG data in daily life with a wrist-worn device and noted that the signal quality was

only high in 50.9% of the time and it varied between activities, with sleep and sedentary activities having higher quality and exercise and activities involving hand movements having particularly low quality.

There is no perfect solution for addressing this problem but there has been research in this area with some results presented in Section 3.1.4.

2.2 SpO₂ Calculation

The PPG measurement of each wavelength allows us to calculate the contribution of the arterial blood to the total absorption of light. As mentioned above, the PPG signal consists of a signal with a DC component and an AC component, where the first is comprised of all the light absorbing components that remain constant. By isolating the AC component of the PPG, by normalizing with their DC component, it is possible to calculate the ratio of ratios, R , described as:

$$R = \frac{AC_{\lambda_1}/DC_{\lambda_1}}{AC_{\lambda_2}/DC_{\lambda_2}} \quad (2.4)$$

Where AC and DC corresponds to the pulsatile and non-pulsatile components of the signal, and λ_1 and λ_2 refers to both wavelengths. The relation between R and SpO₂ can be seen in equation 2.5 where ϵ represents the different extinction coefficients [3].

$$SpO_2 = \frac{R \cdot \epsilon_{Hb\lambda_2} - \epsilon_{Hb\lambda_1}}{R \cdot [\epsilon_{Hb\lambda_2} - \epsilon_{HbO_2\lambda_2}] + \epsilon_{HbO_2\lambda_2} - \epsilon_{Hb\lambda_1}} \times 100\% \quad (2.5)$$

This relation has a number of inherent issues, such as: (1) the extinction coefficients are dependent on the wavelength of the light, as shown in Figure 2.3, consequently any shift in the light wavelength results in a error on the SpO₂. There is also not a perfect agreement of the extinction coefficients across multiple studies [34]. (2) Equation 2.5 assumes that optical path lengths are the same for the different wavelengths. Recent studies have shown that this assumption is not true [3].

In order to overcome these issues, the SpO₂ is empirically determined during the development of the device using a calibration curve, for example, a linear function:

$$SpO_2 = \alpha \cdot R + \beta \quad (2.6)$$

The calibration coefficients, α and β , are specific for each device and are determined by measuring R in several healthy individuals, while measuring SaO₂ in extracted arterial blood or by using a calibrated pulse oximeter as reference [20, 40]. While the full range of the SpO₂- R is not linear, it is possible to approximate using a linear function with low impact on the results [40], especially in the healthy range of values, as seen in Figure 2.5.

There are also some studies using machine learning for calibration-free SpO₂ calculations. This methods are discussed more in depth in section 3.2.2.

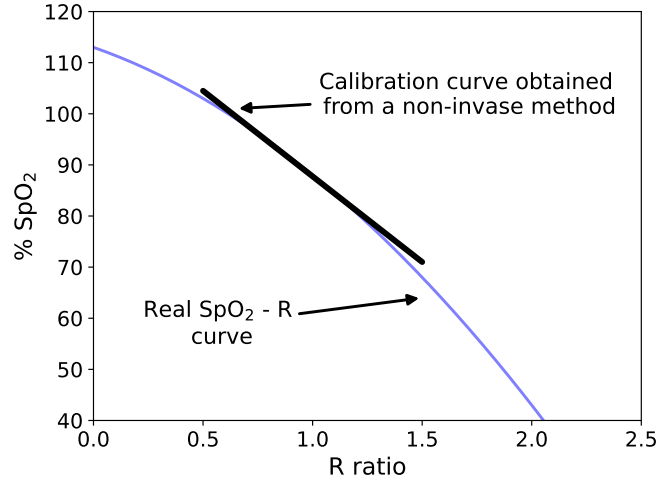


Figure 2.5: Example of a R-SpO₂ calibration curve, adapted from [40].

2.3 Machine Learning

ML is one of the fastest growing areas of computer science, with countless real world applications [41]. ML can be defined as the process of firstly gathering data and then using it to automatically generate an algorithm that can evaluate data that has never seen before, without any help from a human [42].

This development, or learning, can be (1) supervised, if different classes in the training data are labeled, (2) unsupervised, if all data are unlabeled, or (3) semi-supervised, if there are both labeled and unlabeled data in the dataset [41]. This work will be focused on supervised learning.

When developing a ML model, it is usual to transform the raw data into a set of features that can be processed, while preserving the original information. A feature is a property of the data, such as the signal mean, standard deviation, peak-to-peak distance, or more complex and personalized features, designed by the developer. The base for a good machine learning project is its features [43, 44]. They are useful for describing the signal in simpler way, in order to distinguish and characterize different classes.

The supervised learning process starts with a labeled dataset, which means that the data is a pair input, output. Input can be anything, such as pictures, text or sensor measurements. The outputs are real numbers (regression), e.g., SpO₂ calculation, as mentioned in Section 3.2.2, or labels for specific classes (classification), e.g., dog, cat, jumping.

In this work, for an automatic signal quality evaluation, two concepts were thought of: a classification approach and a regression approach. There are several alternatives for ML models, both for classification and regression. The ones tested initially are all supervised learning algorithms and are presented in Table 2.1.

For the classification and regression models, a Random Forest (RF) algorithm was

Table 2.1: Models tested for classification and regression of the data quality.

Classification	Regression
Nearest Neighbors Classification (KNNC)	Nearest Neighbors Regression (KNNR)
Decision Tree Classification (DTC)	Decision Tree Regression (DTR)
Random Forest Classification (RFC)	Random Forest Regression (RFR)
Support Vector Classification (SVC)	Support Vector Regression (SVR)
Gaussian Naive Bayes (GNB)	Ridge Regression
Multi-Layer Perceptron (MLP)	Lasso Regression
	Gradient Boosting Regression (GBR)

selected and thus only this method will be described. Random Forests are built by joining the fundamental principles of decision trees and bagging [45]. The "forest" in the algorithm name refers to the multiple decision trees created, each trained with a random sample of data and features (bagging).

2.3.1 Decision Trees

The basic decision tree models work by dividing the data into two smaller (non-overlapping) sets with a common aspect, using a set of splitting rules [45]. This *divide-and-conquer* algorithm can produce a simple model that is very easy to visualize with a diagram, as can be seen in Figure 2.6. Often, machine learning models can be treated as *black boxes* due to their complex inner workings. Interpretable machine learning is an emerging field, especially in a clinical environment, as the lack of interpretability is one of the major handicaps in its application [46].

The data can also be divided into more than two subsets, called multiway splits. Although this method has been reported to improve readability [47], this feature is not present in the sklearn package [48] for Python, hence only binary divisions will be used.

A basic decision tree divides the data into groups with similar response to a certain feature. This response is a yes-or-no question, forming binary partitions. This is done recursively until a stopping criteria is satisfied. Every group in a decision tree is called a node. The first one, containing all data, is called the root node. The final groups are called terminal nodes or leaves. And every middle group is called an internal node. The connections between nodes are called branches. This structure can be seen in Figure 2.6. For a classification problem, the output is the class with more representation in that leaf. For regression problems, the output is the average of all the samples that fall in that leaf.

If the tree is excessively complex, there is a tendency to overfit [45]. The two most common approaches for limiting the size of a decision tree are: restricting the tree depth and restricting the minimum node size by restricting the minimum number of data points allowed on a terminal node or by restricting the minimum number of data points required for splitting an internal node. When restricting the depth of the tree, the algorithm stops when a certain depth is achieved. A shallower tree is less prone to overfit but may not be

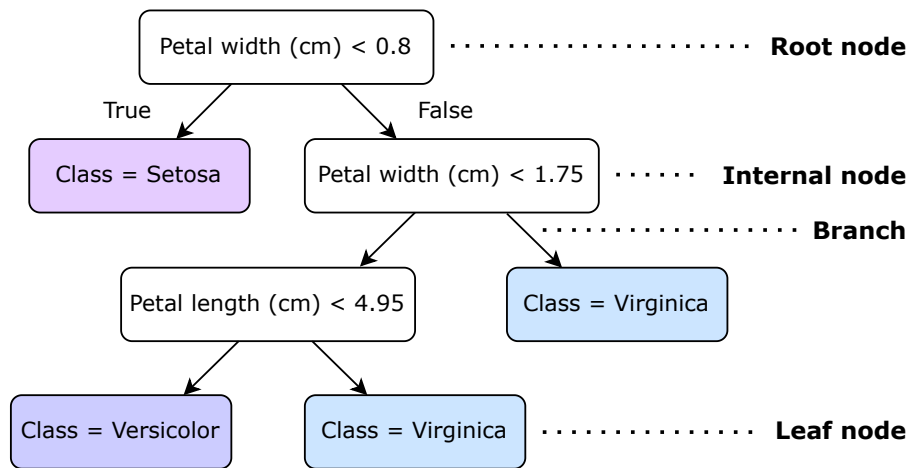


Figure 2.6: Decision tree example with the correct terminology, using the Iris dataset. The tree was limited to a maximum depth of 3 and a minimum of 50 samples to divide a internal node. This was done in order to improve readability, one of the major drawbacks in clinical context.

able to capture the interactions and patterns of complex data [45]. When restricting the minimum node size, the algorithm may not divide internal nodes with less than a certain amount of data, e.g., leaf node must contain more than 10 observations. If there is no limit, the tree has a high variance but poor generalizability, but if the limit is too high, the tree may not capture the interactions between the different features [45]. Figure 2.6 shows a tree with depth restriction and a minimum number of samples required to divide an internal node. The generated tree is very easily read, but its performance can be affected compared to an unrestricted decision tree.

2.3.2 Bagging

Bagging is the process of resampling data by drawing subsets of samples and features with replacement of the original dataset (bootstrapping) and then combining the results achieved by the different models [45]. Even though this method is usually used with decision trees, creating Random Forests, it can be used with any machine learning algorithm. A ML model with bagging follows this steps:

- Create k bootstrap copies of the training data.
- Create k ML models, training each one with a single bootstrap copy.
- Aggregate the decision from the multiple individual learners. In a regression context, the final result is made by averaging the predictions from the different models. For a classification problem, there are two possibilities. The most common approach is to consider the majority vote (the class with the most votes is considered correct) [41]. The other alternative is to average the results of the various learners. The second alternative is used by the scikit-learn package [48], used in this thesis.

Using more trees will not lead to overfitting and, typically, the more trees the better, as the algorithm averages over more high variance individual decision trees [45]. But, eventually, the model performance will stabilize, as can be seen in Figure 2.7. It is important to note that, since multiple models need to be run, both when training and when performing in real-world situations, the more trees present in the forest, higher the computational complexity. This is a limiting factor in wearable technology. Bagging may also remove the ability to easily visualize models that previously were interpretable without difficulty.

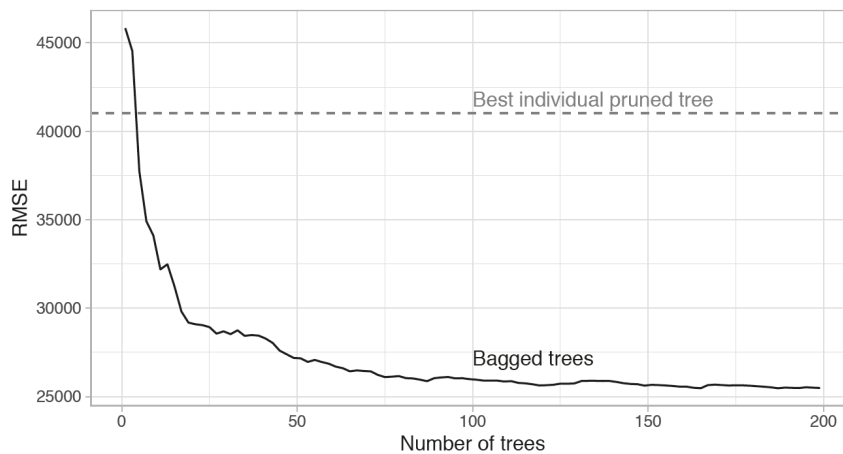


Figure 2.7: Error curve for bagging 1 to 200 decision trees, compared with a single decision tree. It is possible to see that most of the performance increase occurred within the first half of the curve. Reprinted from [45].

2.3.3 Train-Test Sets

The objective of supervised machine learning is to develop a model to predict the results of new data based on the patterns of previous observations. Therefore, there is a need to evaluate the model on its abstraction capabilities. As a result, evaluating the models using the same data as was used for training is not a good indicator for future performance, and the results will be optimistic [49]. Therefore, there is a need to divide the data into two distinct groups:

- Training data: Data used by one or more learning models to develop the classifiers;
- Test data: Data used to calculate the final performance of the model.

The two datasets must be independent and, to an extent, as large as possible in order to obtain the best and most reliable results. Once the models are evaluated the two data blocks can be added in order to produce a final classifier for real application [49].

When working with Big Data there is no problem to perform this division since there is a large sample for both training and testing. The problem comes when there

is a limited amount of data, e.g., the dataset has to be manually annotated or there is limited time for data acquisition. For this type of datasets, a certain amount is reserved for training purposes, this is called holdout process. There is also the need to make sure that the samples for both training and testing are representative of the real world, by representing each class proportionally in both datasets. This procedure is called stratification. Consequently, a general good practice when working with limited datasets is the use of a stratified holdout process to divide the original data.

2.3.4 Performance Evaluation

An important step in the development of a machine learning model is its evaluation. There are two different errors: (1) training error or resubstitution error, and (2) test error. The first evaluates the adaptation to the training set and the latter estimates the capability of generalization, the true prediction error [50].

- **Single Holdout Random Sampling**

This is one of the simplest methods for sampling data. This method randomly selects cases for the train and test datasets. Generally, the test set contains 10% to 30% of the available data, and the training set keeps the remaining data [50].

- **k-fold Random Subsampling**

In this method, the single holdout sampling is applied k times. The learning process is implemented in each training set and evaluated with the corresponding test set. Finally, the performance is estimated by the mean over all k subsets. As with the single holdout sampling method, the train and test set do not overlap, but different train-test sets can overlap. An example of a 3-fold random subsampling can be observed in Figure 2.8(a).

- **k-fold Cross-Validation**

k-fold cross-validation is similar to k-fold random subsampling, with the main difference being that two different test sets do not overlap, which can be seen in Figure 2.8(b). The main dataset is divided into k subsets, or folds, all with approximately the same size and, if desired, with stratification. The final performance is the average for every fold, i.e., the cross-validation accuracy is the mean of all k accuracies. For real-world applications, this is one of the most used methods for estimate the prediction error and the most used number of folds is 10 [50].

For a more accurate error estimate, there is also the possibility to repeat the k-fold a number of times with the same learning method and dataset, but with different subsets, resulting from the random variation in choosing the folds. For example, a standard procedure is to repeat a 10-fold cross-validation 10 times. This involves the learning of 100 different models, which can be time and computational expensive [49].

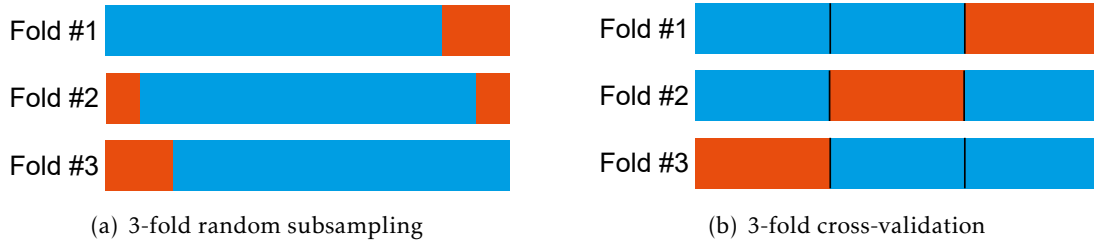


Figure 2.8: 3-fold random subsampling (a) and 3-fold cross validation (b) where the train set is represented in blue and the test set in orange. In the k -fold random subsampling, the number of folds is independent from the ratio of train-test.

- **Leave-One-Out Cross-Validation**

In k -fold cross-validation, when $k = n$, where n is the number of samples, there is a special case and it is called leave-one-out cross-validation. In this process, each model is trained with all samples, except one. Then, the model is judged by its success, or failure, on the single test instance. The results of all n judgments are averaged to calculate the final error estimate. The computational power in this process is elevated, specially when there is a large number of samples. Therefore, this process is mainly used when there is a very small dataset.

- **Subject-Wise Cross-Validation**

In clinical predictions, the goal of a ML model is to generalize to new subjects, and the correct method to evaluate this is to use a slight modification of the standard cross-validation approach, subject-wise cross-validation [51]. With this technique, the training and test folds contain data from different subjects, therefore, the results are more generalizable to real-world scenarios, as represented in Figure 2.9. Common cross-validation can still be used but the accuracy of the model can be very overestimated [51].

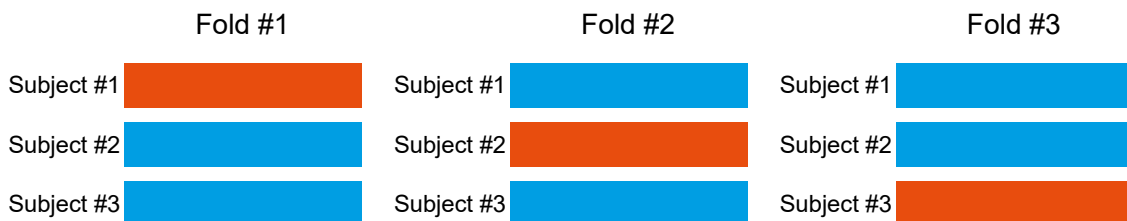


Figure 2.9: Subject-wise cross-validation example where the train set is represented in blue and the test set in orange.

2.3.4.1 Classification Metrics

As different metrics allows for assessing different information, there are several metrics that can be used to evaluate the different models developed.

A confusion matrix is a table that compares the results obtained with the expected outcomes, in which the rows represent the actual class and the columns represent the predicted classes, as seen in Table 2.2. For binary classification, the confusion matrix can be easily represented and each cell represents a specific metric. The true positive (TP) and true negative (TN) represent the correctly classified instances, for the positive and negative classes, respectively. The false positive (FP) and false negative (FN) denote misclassified instances. The FP case represent a negative instance classified as positive class, and the FN case represent opposite. For multiclass classification a confusion matrix can still be created but these parameters cannot be directly extracted from it.

Table 2.2: Confusion matrix for binary classification.

	Predicted Positive Class	Predicted Negative Class
Real Positive Class	True Positive (TP)	False Negative (FN)
Real Negative Class	False Positive (FP)	True Negative (TN)

Other metrics chosen for the evaluation of the classification models were: accuracy, precision, recall, F-score (also known as F1 score) and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). A brief explanation appears bellow. Note that the following expressions are for binary classification but, with some adjustments, they can be used for multiclass classification.

- **Accuracy:** Evaluates the ratio between the correct predictions over the total number of samples evaluated [52].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2.7)$$

- **Precision:** Evaluates the positive samples that are correctly evaluated from the total predicted samples in a positive class [52].

$$Precision = \frac{TP}{TP + FP} \quad (2.8)$$

- **Recall:** Evaluates the ratio between the positive samples that are correctly classified and the total of rightly classified samples [52].

$$Recall = \frac{TP}{TP + FN} \quad (2.9)$$

- **F-score:** Represents the harmonic mean of precision and recall, balancing the information from both metrics [52].

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2.10)$$

- **AUC-ROC:** This is one of the most popular ranking type metrics [52]. This metric reflects the model performance better than the accuracy [52].

$$AUC - ROC = \frac{S_p - n_p(n_n + 1)/2}{n_p n_n} \quad (2.11)$$

Where S_p represents the sum of all positive examples ranked and n_p and n_n stand for the number of positive and negative examples, respectively.

2.3.4.2 Regression Metrics

Unlike with classification, in a regression problem it is unlikely to predict the exact real values, as it deals with continuous variables, thus infinite values. The question is rather how close the model prediction is to the real value. The metrics chosen were R^2 , root mean square error (RMSE), and mean absolute error (MAE).

- **Mean absolute error (MAE):** Is simplest form of estimating the performance of a regression model, by subtracting the predicted values from its corresponding true value [53].

$$MAE = \frac{1}{N_{samples}} \sum_i^{N_{samples}} |y_i - p_i| \quad (2.12)$$

Where y_i represents the real output and p_i represents the predicted output.

- **Root mean square error (RMSE):** Is the square root of the mean square error (MSE). This metric, unlike the MAE which treats all errors the same, penalizes larger errors [53]. This metric is also the metric implemented in the pulse oximeters that follow the ISO 80601-2-61 directives [54].

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N_{samples}} \sum_i^{N_{samples}} (y_i - p_i)^2} \quad (2.13)$$

- **R^2 :** It is the square of the correlation coefficient (R) and is a good measure to evaluate how well a model fits the dependent variable, where a R^2 value close to 1 indicates a strong correlation [53]. R^2 can also be negative. This happens when the data as a correlation worse than a horizontal line [55].

$$R^2 = 1 - \frac{\sum_i^{N_{samples}} (y_i - p_i)^2}{\sum_i^{N_{samples}} (y_i - \bar{y})^2} \quad (2.14)$$

Where $\bar{y}$ represents the real output average.

2.3.5 Under and Overfitting

There is an important tradeoff in machine learning with underfitting and overfitting the model to the training dataset. Underfitting occurs when a model fails to predict the results on the training data [43]. Overfitting is the opposite, the model adapts too much to the training dataset performing very well for the training data but poorly for the testing data, preventing the abstraction needed. This is a central problem in ML, with more papers written about it than about any other topic [56]. Both cases are represented in Figure 2.10 and possible solutions are:

- Underfitting: Using a more complex model; Add more training data; Construct new features with more predictive power.
- Overfitting: Using a simpler model; Decrease the ratio features/samples.

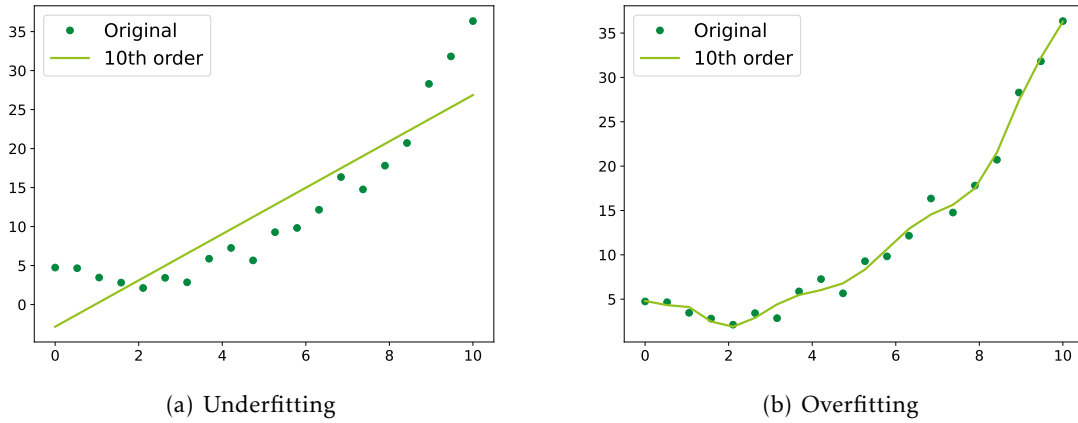


Figure 2.10: Underfitting example (Linear model) and overfitting example (10th order model) of a known function.

2.3.6 Balancing Data

An imbalanced dataset has a non-uniform distribution across the multiple classes. This distribution can vary from a slight bias to a severe imbalance, where one class can have thousands less examples than the other classes. A slight imbalance is often not a problem and can be treated like a normal classification problem, but many ML models assume a balanced class distribution and can have a poor performance with highly unbalanced datasets [57].

One of the solutions is data sampling. There isn't a perfect data sampling method [57] and, for this work, the random undersampling method was used. This is the simplest undersampling method. With this approach a random sample of instances in the majority class is deleted.

State of the Art

3.1 Photoplethysmography

The PPG signal is highly affected by several factors, such as: the LEDs wavelength and number used, as well as their positioning, the quality of the signal and its processing. These factors will be analyzed in the following sections.

3.1.1 LED Count and Wavelengths

As can be seen in Table 3.1, there are multiple alternatives to the LEDs quantity and wavelength. The most common combination is red-infrared, but the selection is often based on personal preferences of the researchers [58]. Hossain, Kwon, and Kim in [58], showed that the red-green pair should give more accurate data. Likewise, Banik et al. in [59], states that using a green LED is more important than using a IR LED. This is because green light is absorbed closer to the skin surface, which improves the signal-to-noise ratio by having a shorter path length, that results in less attenuation [60]. This is specially helpful in reflectance pulse oximeters. It has also been reported that green light provides signals with improved resilience to motion artifacts [23].

Table 3.1: LEDs used and their mode of operation in SpO₂ devices, where IR represents an infrared LED.

Number of LEDs	Wavelength	Mode of operation	Reference
3	Green(525 nm), Red(660 nm) and IR(850 nm)	Reflectance	[15]
8	Green(525 nm) e Orange(595 nm)	Reflectance	[23]
2	IR(761 and 818 nm)	Transmission	[28]
1	Red(680 nm)	Transmission	[29]
2	Red and IR	Transmission	[30]
2	Red(620-625 nm) and Green(520-525 nm)	Reflectance	[59]
2	Red(660 nm) and IR(890 nm)	Reflectance	[61]

It is reported that increasing the number of transmitted wavelengths may improve the accuracy of the measurements, as well as giving the ability to monitor other types of hemoglobin [62], as shown in Section 3.2.3. Shen, Chen, and Wu [62] used a xenon

lamp in order to obtain a light with a wavelength range from 400 nm to 1100 nm and two six-channel sensors for acquiring twelve wavelengths. These values were fed into a deep learning network, which automatically determined the SpO_2 value. They showed that their multi-wavelength spectral measurement had a low error in eliminating the effect of the tissue and had improved the SpO_2 measurement accuracy. There is also a commercially available pulse oximeter that uses up to 8 wavelengths, Massimo SET pulse co-oximeters [63, 64].

There is also research by using a single LED associated with a multi-spectral optical sensor [29]. Von Chong et al. [29] used this approach, by using a single LED (680 nm) and a Buried Quad Junction (BQJ) photodetector. This PD provides, for a given wavelength, a specific and unique current combination through its four channels. Since the spectrum from the light transmitted through the finger also depends on the extinction coefficient of the whole blood, a change in oxygenation level results in a different incident light spectrum. This allows for a signature distribution of BQJ's currents for each SpO_2 value. The authors showed, from three in-vivo results, that the SpO_2 estimations fit into the 4% tolerance range. They also claim that precision and robustness could be improved by using known algorithms of data processing, and that the device could also be used to estimate other types of hemoglobin.

3.1.2 Sensor Location

The most common placement for transmission pulse oximeters is the finger, often the forefinger [3]. This was used in the first pulse oximeter prototype [22] and it is a good placement considering the anatomical structure of the finger [3]. This placement highly depends on the peripheral blood flow, hence situations such as peripheral vasoconstriction, hypovolemia and cold fingers can provide false readings [3]. This limitation inspired researchers to develop solutions utilizing other measurement sites, such as forehead, wrist, finger (in a ring configuration), in-ear, earlobe, nasal, sternum and tongue pulse oximetry [3, 31, 65].

3.1.2.1 Wrist Pulse Oximetry

The main advantage of a wrist-worn device is the ability to wear throughout the day with minimal discomfort, since it can be easily worn in a form of a band or watch. Many wrist PPG sensors have become available in the last years but most of them can only be used for personal monitoring and are not practiced clinically [66]. An example of a clinical accurate pulse oximeter is the BORA Band developed by Biosency [67].

In order to utilize their principal advantage, there are some key features to take into consideration [39]: (1) devices that are waterproof and require less charging are more likely to be used by longer periods of time; (2) data should be stored on the device or streamed via a reliable connection to a secondary device; (3) the device should be sufficiently comfortable for extended use.

This type of devices have some disadvantages: They are highly susceptible to motion artifact [17] and a slight position change of the measurement site can affect the device performance [17, 66]. However, there are cases in which the artifacts can be dealt with, overcoming this limitation.

3.1.3 Quality Assessment

As mentioned in Section 2.1.3, the PPG signal is highly affected by a number of factors. In a supervised setting, such as the hospital, if a signal has too poor quality an operator can intervene and the recording is repeated [68]. This step is impractical in a remote setting, the case of a wearable device. Therefore a PPG automatic signal quality assessment phase on a wearable device is beneficial.

Various techniques have been successfully implemented. Most methods can be categorized under beat segmentation [69–71] (where each PPG pulse is represented as a straight line originating at the pulse start and ending at its primary peak) and non beat segmentation [18, 72–76] (where the evaluation is made using time windows, that can have different sizes). The non beat segmentation approach will be analyzed, as this technique is adopted by the majority of the works [18] and was used in this work.

Most of the current research uses binary classification [18, 73–75]. This may not be the most correct approach as a lightly corrupted PPG signal can still provide information, such as the heart rate [18]. As an alternative, Prasun, Mukhopadhyay, and Gupta [18] classified the signals into three different classes ('clean', 'partially clean' and 'corrupt', illustrated in Figure 3.1), Elgendi [76] also divided between three qualities ('Excellent', 'Acceptable' and 'Unfit') and Pradhan, Rajan, and Adler [72] differentiated between five qualities (1 to 5, with 1 representing the lowest quality and 5 representing the highest quality). There is also research for quality classification in the presence of abnormal HR, such as atrial fibrillation [74, 75].

3.1.4 Artifacts Removal

Due to the PPG sensitivity to motion artifacts, one of the most researched topics in PPG enhancement is MA removal [77]. Standard filtering techniques, despite still being used, cannot deal with all types of artifacts, as they act over a fixed window of frequencies. Hence this technique can only be used when the artifacts and signal frequencies band do not overlap. Two important alternatives are independent component analysis (ICA) and adaptive noise cancellation (ANC).

As mentioned in 2.1.3.1, signal can be decomposed in two main components, a component representing the MA and a component representing the vascular dynamics, the interest signal. This is the principle behind ICA, which can decompose a signal into its sub-components, assuming they are independent of each other [78]. This method requires the use of multiple PPG sensors and is reported that the artifact-free signal is not always restored [79], but new methods are being developed with promising results [80].

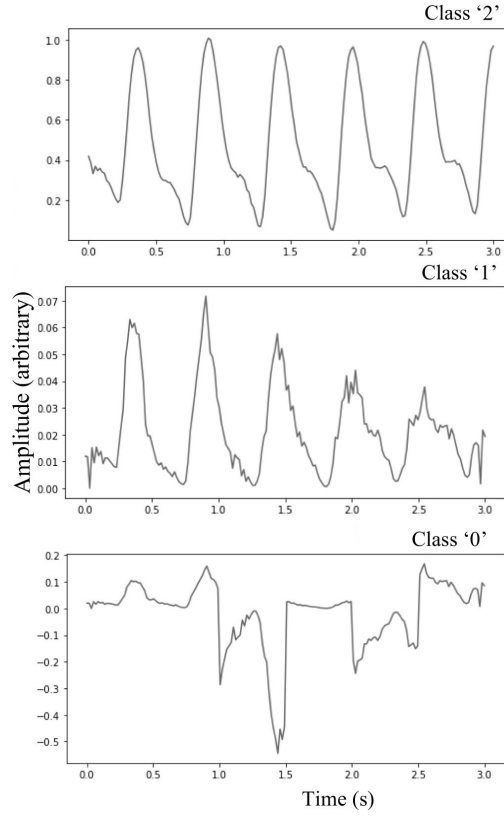


Figure 3.1: Sample PPG segments from three different classes, reprinted from [18]. The class '0', '1' and '2' represents the 'corrupt', 'partially clean' and 'clean', respectively.

ANC is also a very useful technique when noise references are available. For motion artifacts, an accelerometer can be used as the source of the motion information. As depicted in Figure 3.2, an ANC algorithm obtains the desired signal by subtracting the accelerometer signal from the PPG signal [81]. The noise signal cannot be directly subtracted from the PPG signal, it needs to pass firstly by an adaptive filter, consisting of a fixed digital filter and an adaptive algorithm with changing coefficients. There are several adaptive algorithms such as least mean square (LMS), normalized LMS and fast transversal least square (RLS) [79].

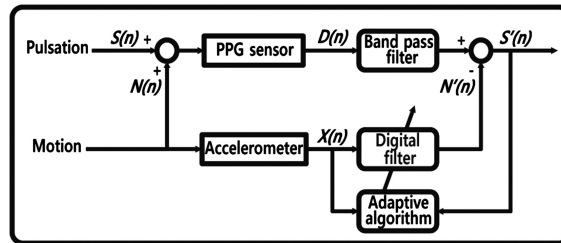


Figure 3.2: Block diagram of an adaptive noise canceling algorithm. Reprinted from [81].

Due to being a simple algorithm, ANC methods are fast [33] and their main disadvantage is the need for extra sensors.

3.2 Clinical Applications

Due to the rise of wearable technologies, thus availability of large amounts of raw data, and the increase in computational power, there has been an increased interest in implementing machine learning in PPG devices. There are two possibilities with ML algorithms: (1) classification, such as detection of signal quality [72, 74, 75, 82] or whether a patient has or not coronary artery disease or atrial fibrillation [74]; (2) parameter assessment, where a parameter is estimated, e.g., blood pressure [83–89], SpO₂ [17, 90, 91], COHb and RR [69, 70, 92–94].

3.2.1 Blood Pressure

The cuff-based sphygmomanometer is the most used method for blood pressure assessment [83]. However, BP varies continuously while this method can only measure BP values intermittently. There are 24-h ambulatory BP monitoring devices that can assess blood pressure during the day, but are usually cuff-based methods that can be distracting and uncomfortable [83, 85]. Hence, there is a need for a cuff-less, non-invasive, wearable, long-term BP monitor.

Although the relationship between blood pressure and PPG waveform is not totally understood, there are studies in this area with cuff-less measurements of BP [84–87]. Many of these studies are based on healthy individuals and artifact-free signals [87–89]. Therefore the signal quality is optimal and can perform poorly in subjects with cardiovascular diseases [85].

3.2.2 SpO₂

The traditional algorithm for SpO₂ estimation uses the ratio R, with its expression presented in Equation 2.4, and a set of volunteers. In a calibration phase, a linear regression is used to model the relationship between the R ratio and the real SpO₂ valued, obtained using a previously calibrated pulse oximeter [91, 95]. Polynomial regression can also be used [17, 59]. This calibration method can be susceptible to ambient light [90] and depends on the volunteers characteristics, such as, skin color and age [95]. Even though the calibration process involves inducing a light hypoxemia, which can be difficult and dangerous, this methods could also be improved with more data, especially for lower values of SpO₂ (<75%) [40, 91].

There is also some studies using ML models to extract SpO₂ values from the PPG signal. Priem et al. [17] used a deep neural network, with the inputs being 10 s windows of three channels (green, red and infrared). These windows were represented by the AC component, the filtered PPG signal, and the DC component, the average of each channel. They exceeded clinical requirements, despite lacking some improvements, such as: assessing the computational cost of the model and evaluating the algorithm on signals acquired during movement. Venkat et al. [90] used various features extracted from the

AC and DC components of the red and IR channels in order to train several ML models. The bagged tree classifier achieved an accuracy of 96%. Despite the good results, the SpO₂ values recorded were skewed in favor of higher values, which led to an increase error on the 81%-85% range.

Therefore, when developing a clinical pulse oximeter, several aspects are needed to be taking into account, such as: the conditions on which the calibration data was acquired and the SpO₂ values used in the calibration process.

3.2.3 Carboxyhemoglobin Saturation

COHb is formed when carbon monoxide (CO) binds with hemoglobin. CO binds in the same sites as oxygen (O₂), but with an affinity 200-250 times greater [96]. This results in a competition between O₂ and CO, which can rapidly cause a functional anemia [96]. For this reason there was a need to develop a device capable of detecting the concentrations of COHb. Rainbow-SET Rad-57 Pulse CO-Oximeter (Masimo Inc., Irvine, CA) is a pulse oximeter that uses eight wavelengths, rather than the two wavelengths usually employed. It is approved by the US Food and Drug Administration for providing information of SpCO (COHb saturation) and SpMet (MetHb saturation), along with the conventional SpO₂ value [63].

3.2.4 Respiratory Rate

The RR can also provide information about hypoxemia as the body attempts to correct it by increasing the breathing frequency [97]. This measurement provides information that pulse oximetry cannot offer but is often not recorded, even among patients with respiratory problems [97].

RR affects the PPG signal in three ways [69, 70, 92–94], called respiratory-induced variations (RIV), illustrated in Figure 3.3:

Respiratory-induced intensity variation (RIIV): The variations in the intrathoracic pressure, leads to a change in the signal baseline. This is due to the fact that, during the inspiration, the intrathoracic pressure decreases resulting in a decrease in the venous pressure, increasing venous return [94]. In the expiration the opposite occurs.

Respiratory-induced frequency variation (RIFV): The respiration causes a variation in the heart rate, increasing it during inspiration and decreasing it during expiration. This causes a variation on the pulse frequency. This variation is also referred as respiratory sinus arrhythmia [93].

Respiratory-induced amplitude variation (RIAV): During the inspiration there is a decrease in the cardiac output, due to reduced ventricular filling [69]. This will result in a decrease of the pulsatile component of the signal.

Window sizes for calculating these RIV can vary between 16 to 128 seconds [70]. Smaller window sizes yield a shorter computation time and provide higher time resolution, while larger windows are reported to provide better accuracy and decrease the lowest

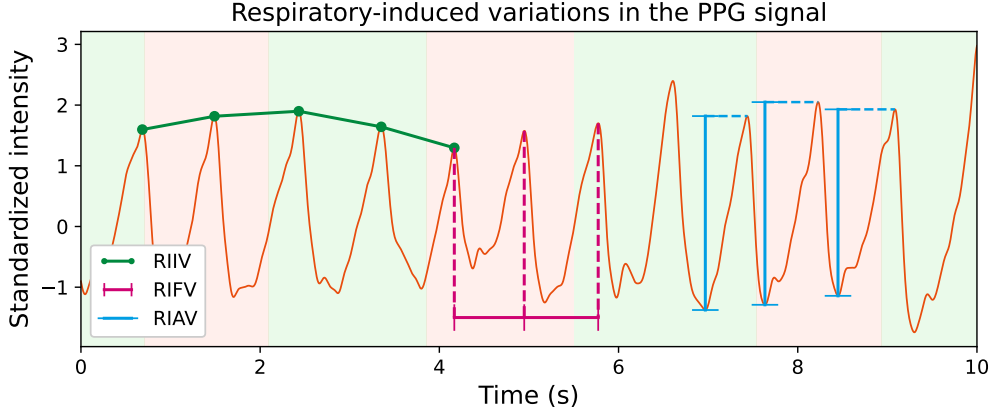


Figure 3.3: Respiratory-induced variations on the PPG signal. RIIV is the baseline variation, RIFV is the respiration effect on the HR and RIAV is the change in pulse strength. The background color represents the respiration phases, where the inspiration is represented in green and the expiration is represented in red.

detectable RR (e.g., for a 32s window the minimum detectable RR is 3.75 breaths/min (brpm)) [69, 70].

These three variations can then be used to estimate the RR. There are multiple methods that can be applied for this estimate, such as based on wavelets [94] and frequency responses [69, 70, 92, 93]. Only the latter will be analyzed.

Karlen et al. [69] calculates the fast Fourier transform (FFT) for every RIV. Then, the frequency with maximum power, within the expected respiratory range (0.067-1.08 Hz or 4-65 brpm), is extracted. The peak frequency from the three RIV are combined by calculating their mean, resulting in the RR. The authors also implemented in the algorithm that only reliable estimations are selected. These estimates have a standard deviation lower than 4 brpm between the three RIV and correspond to windows without artifacts.

Dubey, Constant, and Mankodiya [92] also use the three RIV, along with the spectral kurtosis of the PPG signal. These values are then fed into a Extreme Learning Machine (ELM) that can learn complex relationships for the RR estimation.

Pulse oximetry is considered a indicator of respiratory problems with delay. Therefore, a combination of this method allied with an accurate assessment of the RR, in a single sensor, can provide an earlier indication of respiratory complications [94].

3.2.5 Heart Rate

Heart rate is one of the the main vital signs [97]. Despite the HR information being accurately measured at rest with a PPG device, the performance drops when motion artifacts are present, such as in daily activities or exercise [6, 79]. The current gold standard reference method for HR assessment is the electrocardiogram (ECG) [6], which is more robust to MA than PPG than the PPG [79].

As the periodicity of the PPG signal derives from the heartbeat, the HR can be easily estimated, and the main obstacle are the presence of MAs. Hence most studies approach is to first remove MA from the PPG signal [79, 98].

Currently, there are no specific clinical standards that must be followed for wearable devices [11]. Nevertheless, physicians tend to accept wearable devices measurements, therefore, a standardization on the accuracy provided by these may be needed [11].

Materials and Methods

This Chapter presents the materials and methods used in this thesis. Therefore, it comprises the description of the (1) data acquisition protocol used, (2) PPG signal processing, including signal quality assessment, (3) feature extraction and selection, (4) machine learning models used and (5) physiological metrics extraction.

4.1 Data Acquisition

In this work, two protocols were developed. The initial protocol was developed for a quality assessment evaluation and the main goal was to simulate real-world situations. The second protocol has a better overall signal quality and was developed to extract the SpO₂ value.

biosignalsplux devices from PLUX Wireless Biosignals S.A. [99] were used. This allows the setup of a system for advanced biosignals research. The hub, represented in Figure 4.1(a), is a 8-port device that collects, digitizes and transmits all signals from the sensors and accessories via Bluetooth to a computer. This arrangement supports the acquisition of signals up to 16-bit resolution and 8000 Hz sampling frequency. In this study the ECG, inductive respiration (RIP), SpO₂ (represented in Figure 4.1) and accelerometer were acquired and the OpenSignals (r)evolution software was used, which is also from PLUX Wireless Biosignals S.A. [99].

Before starting the acquisition, the full procedure was explained to the participants of the study, including which signals would be acquired and the different activities to be performed. After that, all subjects read and signed an informed consent explaining the study. After this, all the required sensors were placed and the OpenSignals (r)evolution software was launched.

An ECG sensor with three electrodes and a RIP sensor were placed according to the Figure 4.2. The SpO₂ and accelerometer sensors were placed on the posterior plane of the wrist where the subject would wear a watch, to mimic a smartwatch placement. The SpO₂ sensor, shown in Figure 4.1(b), has a red and an infrared LED, with wavelengths of 660 nm and 950 nm respectively, and works in reflectance mode. All the sensors were

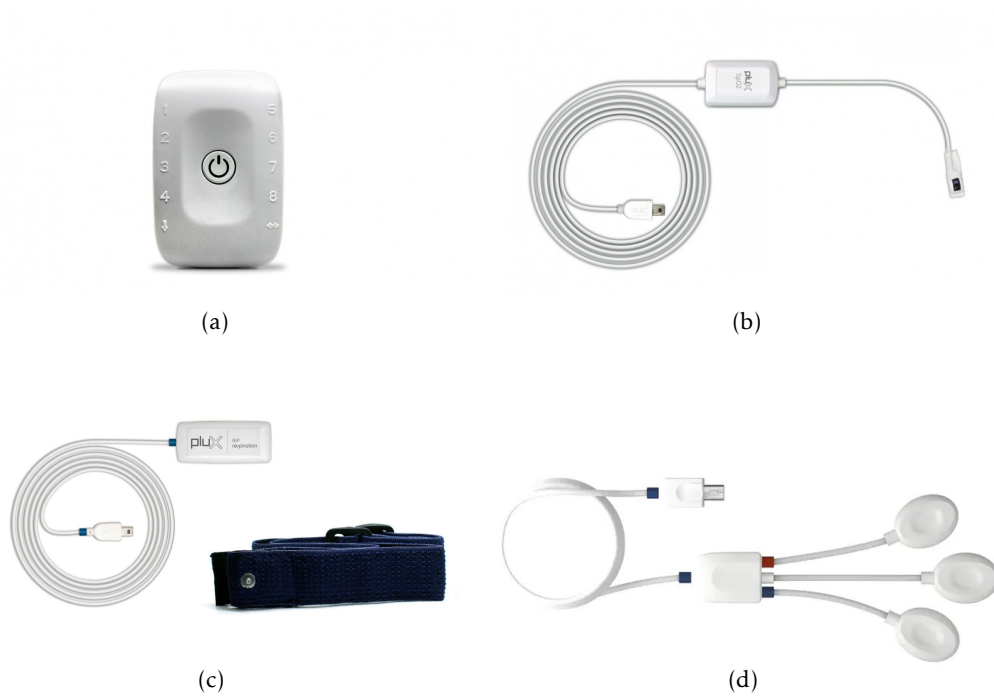


Figure 4.1: biosignalsplux 8-port hub (a), SpO₂ sensor (b), inductive based respiration sensor (c) and ECG sensor (d).

sampled with a frequency of 200 Hz and with 16-bit resolution.

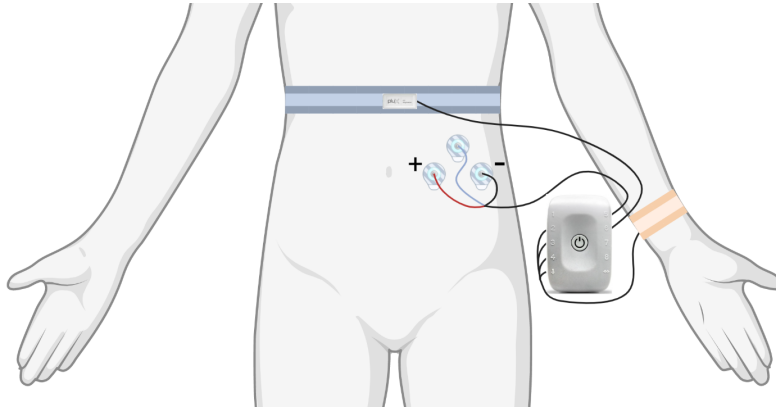


Figure 4.2: Sensors placement. The respiration band was located in the lower thoracic region and the three ECG electrodes were placed below it. The ECG was arranged in an inverted lead I configuration. This was done in order to simplify the electrodes connection while maintaining the R-peaks intact.

Initially, the subjects were asked to fill a survey with personal information, including their dominant hand and where would they wear a watch. This information was used to decide on which wrist the sensors were connected. All information collected in the survey can be seen in Table 4.1. The biosignals obtained during this survey were discarded from further analysis.

Table 4.1: Personal information collected in the survey.

Survey information	
Age	Sex
Smoker? How many?	Coffee in the last 2h?
Dominant hand	Watch wrist
Physical activity level	Exercise in the last 24h?
Medical information	Medication

4.1.1 Acquisition Protocol for Quality Assessment

The data acquisition protocol is represented in Figure 4.3. It was developed to simulate real-world activities, with a two minute pause between every task.

After the initial survey and the two minute pause, a typing task was performed. This task consisted in typing a text for three minutes in a computer. A typing tutor (<https://www.typing.com/>) was used and, once the task was finished, the words-per-minute and accuracy were provided and noted [100]. The writing task consisted on writing by hand a 160 word text, with the option of stopping before the end of the text. The walking task was a simple walk with natural arm movement. Its duration was not defined in the protocol. Additionally, annotations were made using an OpenSignals plugin, marking in the beginning and the end of every activity.

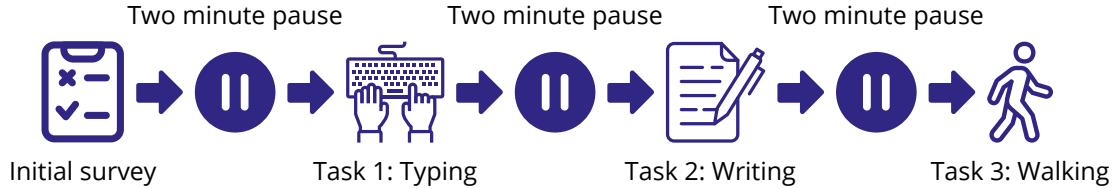


Figure 4.3: Experimental protocol for quality assessment.

4.1.2 Acquisition Protocol for SpO₂ Extraction

In order to estimate the SpO₂ value, a second acquisition protocol was developed, illustrated in Figure 4.4. This protocol was developed with less tasks prone to motion artifacts. Therefore, the three tasks were replaced by two rounds of the Wim Hof Method (WHM) breathing exercises, guided by the instructions provided in the website [101]. The purpose of these exercises is to allow the subject to be in apnea for longer periods of time than in normal circumstances, which causes the SpO₂ to drop [102]. The pause moments were also shortened to one minute each in order to reduce the total acquisition time.

As these acquisitions main purpose was for extraction the SpO₂ information from the PLUX Wireless Biosignals S.A. [99] sensors, an additional sensor was used, a Contec CMS50D+ Pulse Oximeter, represented in Figure 4.5. The Contec sensor is fitted with

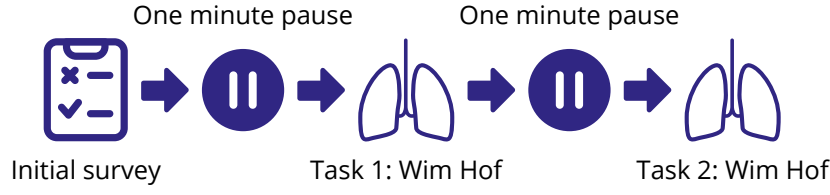


Figure 4.4: Experimental protocol for SpO₂ extraction.

two LEDs, with wavelengths of 660 nm and 880 nm and can measure the SpO₂ with a $\pm 2\%$ absolute error in the 70% to 100% range. It also provides information relative to the pulse rate with an accuracy of ± 2 bpm. This sensor was placed in the non dominant hand, in order to have little movement as possible, hence recording the data with higher accuracy.



Figure 4.5: Contec CMS50D+ Pulse Oximeter [103].

It can also be connected to the biosignalsplux 8-port hub, showed in Figure 4.1(a), and integrated with the OpenSignals (r)evolution software, an essential step to ensure signal synchronization.

4.2 Data Processing

Due to the PPG signal characteristics, especially the AC and DC components, there is a need to process the acquired data before extracting important information. The workflow for the data processing can be seen in Figure 4.6 and it was performed on both channels, red and infrared.

4.2.1 Digital Filter

According to Allen and Kyriacou [3], the PPG bandwidth is up to 5 Hz, and all relevant signal characteristics could be extracted within this frequency range. The signal's DC component can also be removed by using a high-pass filter, with cut-off frequency up to

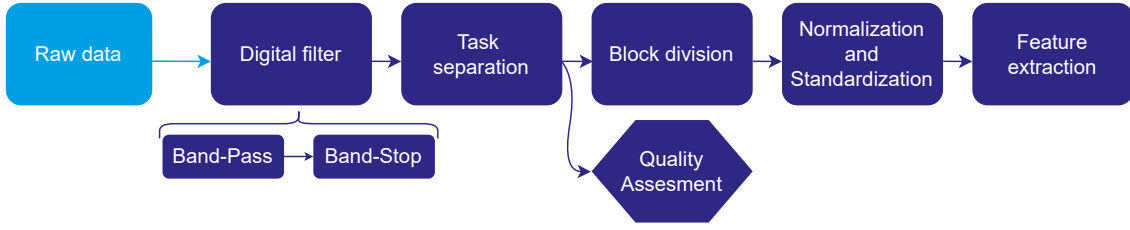


Figure 4.6: Data processing workflow.

0.5 Hz. Therefore, a band-pass filter with cut-off frequencies of 0.1-0.5 Hz and 10-40 Hz can be used.

The digital filter phase, first step shown in Figure 4.6, consists of two different filters. Firstly a 5th order Butterworth band-pass filter with cutoff frequencies of 0.5 Hz and 10 Hz. The purpose of this filter is to remove the DC component and attenuate the high frequency noise present in the signal, while preserving the biosignal information. Then, a 3rd order Butterworth band-stop filter with cutoff frequencies of 40 Hz and 60 Hz was applied. The purpose of this filter is to remove, with greater detail, the 50 Hz noise coming from the grid. Both filters were implemented forward and backward to the signal, resulting in a signal with zero phase and a filter with twice the order [104]. This is an important aspect since signal synchronization with the other sensors is crucial and a shift in phase can result in a misaligned signal. An example of a raw and filtered signal can be seen in Figure 4.7.

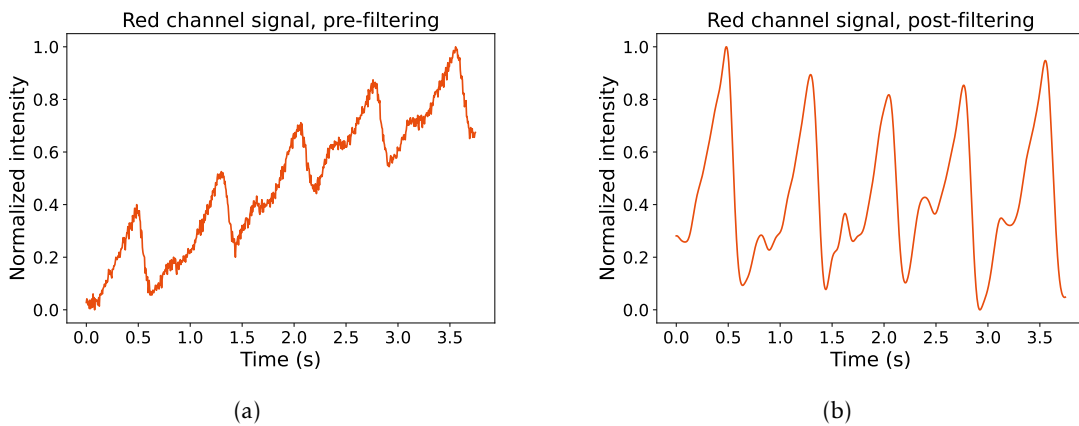


Figure 4.7: Raw (a) and filtered (b) red channel signal example. The filtered signal resulted from a band-pass and a band-stop filter applied forward and backward to the original raw signal.

4.2.2 Task and Block Division

The data was then divided into different tasks, based on annotations made during the acquisition. During this division, the first and last seconds from each activity were discarded. Then, each task was divided into blocks of 3 s each. For real-time analysis, a delay of a couple of seconds is a common aspect [76]. Elgendi [76] analyzed the effect of the window size on the performance of a signal quality index (SQI). The author found that, for that particular SQI, windows with 2 s, 3 s and 5 s were optimal for differentiating between three qualities. Prasun, Mukhopadhyay, and Gupta [18] also used a 3 s window for real-time quality assessment using ML. Therefore a 3 s window was selected for this work. For an analysis closer to real time, a sliding window with 2 s overlap was used, that is, for each block, there are 2 s of data common with the previous block, as shown in Figure 4.8.

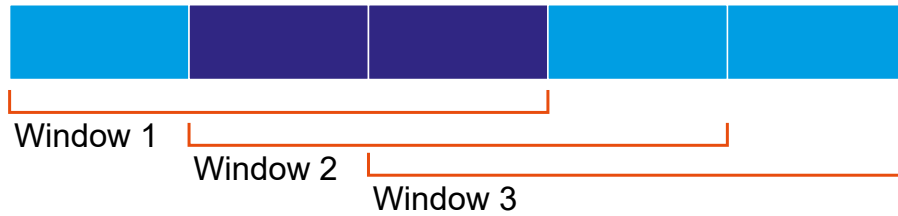


Figure 4.8: Schematic of a sliding window with overlap where each section represents one second. Note that, as an example, the dark blue sections represent the seconds of overlap between the first and second window.

4.2.3 Quality Assessment

As mentioned in Section 2.1.3.1, the photoplethysmography signal is highly susceptible to motion artifact, which may hinder its analysis, therefore a quality assessment stage was developed.

In order to make a supervised ML model for quality assessment, it was necessary to manually evaluate all data beforehand. This evaluation was done on a continual basis, meaning the data was evaluated after filtering and before dividing into different blocks, as shown in Figure 4.6. This allowed for more flexibility when choosing the window size, since, by using this method, all samples have a corresponding quality. This evaluation was carried out by two master students in Biomedical Engineering and an example of each quality is shown in Figure 4.9. For this rating, the ECG signal was also used to help assess where there were expected peaks in the PPG signal. The data was divided into three qualities and they can be described by the following definitions:

Class '2' Optimal quality. There are clear peaks and the signal features are easily identifiable.

Class '1' Suboptimal quality. The signal features cannot be identified, but it is possible to observe the peaks related to the heart rate.

Class '0' Corrupted, i.e., there can be motion artifact, signal loss, or other external influences on the signal quality. No information can be obtained from the signal.

To establish the quality of each block, three interpretations were developed:

'Average' quality The quality of the block is the average quality of every sample in that window. Depending on whether a classification or regression model was being developed, this value was rounded to the nearest integer or not, respectively. An example of this quality interpretation is shown in Figure 4.10.

'Strict Average' quality This interpretation is similar to the 'Average' quality. The only difference is that if the block has more than a third of its quality with class '0', the quality of that block was automatically class '0'.

'Mode' quality The quality of the block is the most common quality in that window, that is, the statistical mode in that window.

4.2.3.1 Inter-annotator Agreement

Since the data was manually evaluated by two different annotators, where each annotator evaluated only a part of the dataset, a measure of the agreement between the two annotators was needed. The Cohen's kappa coefficient (k) was used [105]. If both annotators are in complete agreement then $k = 1$. Otherwise, if the agreement between the annotators is the expected by randomly classifying the data then $k = 0$. The expression for this coefficient can be seen in Equation 4.1.

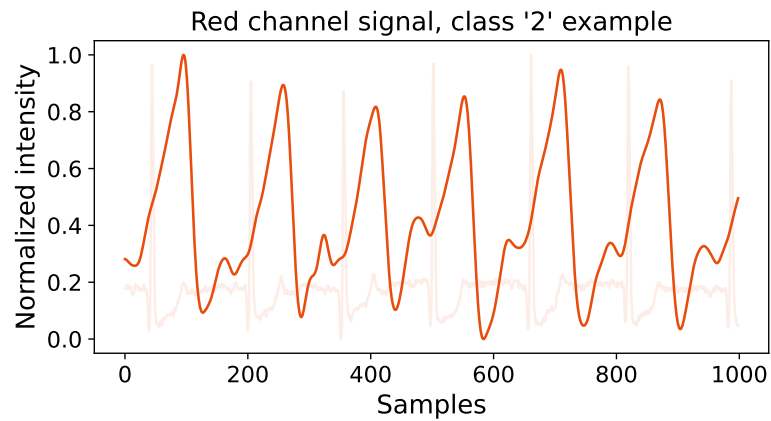
$$k = \frac{p_0 - p_e}{1 - p_e} \quad (4.1)$$

Where p_0 represents the observed agreement among annotators and p_e represents the hypothetical probability of chance agreement, calculated using the expressions 4.2.

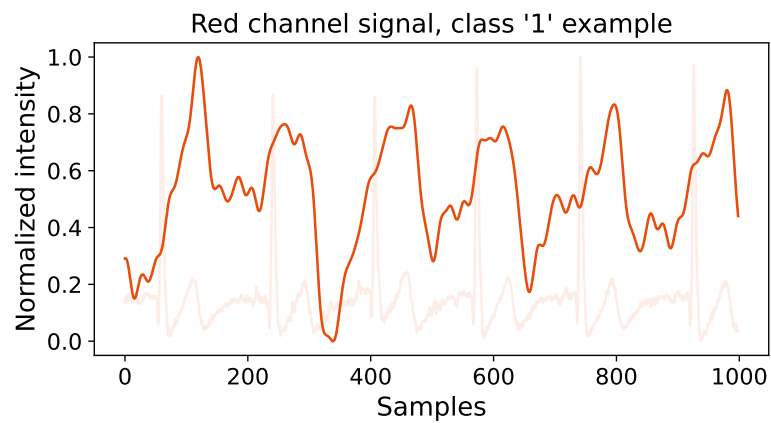
$$p_0 = \frac{N_{agreement}}{N_{total}} \quad p_e = \frac{1}{N_{total}^2} \sum_i n_{iA} n_{iB} \quad (4.2)$$

Where i represents the number of categories, in this case the three qualities, and n_{iA} and n_{iB} represent the number of times that the annotator A and B predicted the category i , respectively.

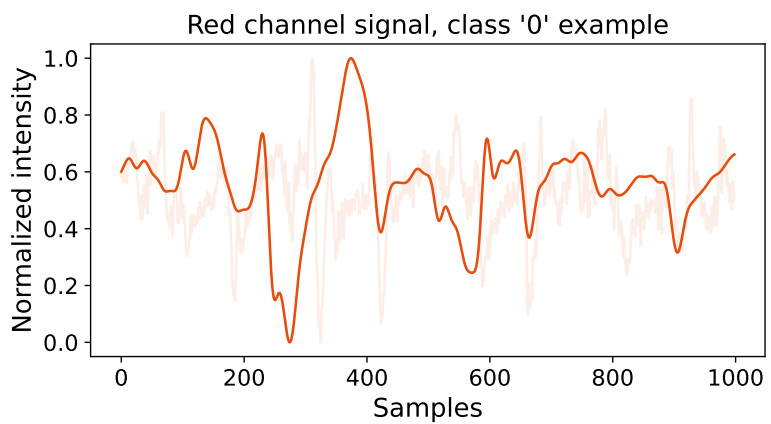
The ideal situation is $k = 1$. In this case both annotators are in perfect agreement, indicating a more precise classification, which improves the data used for ML development.



(a)



(b)



(c)

Figure 4.9: Data quality examples. Data with quality '2' has clearly visible peaks with the signal features still present. Data with quality '1' has a visible hearth rate but some peaks can have plateau and the amplitude varies between peaks. Data with quality '0' has no visible peaks representing the hearth rate or any other features. It is also possible to see the ECG signal, used for the manual evaluation. In the case of class '0', ECG signal also shows high noise and artifacts.

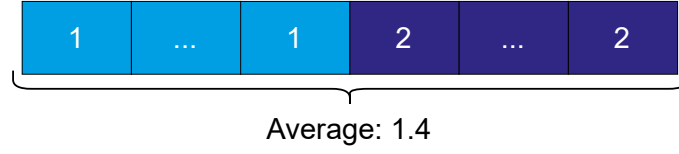


Figure 4.10: Example of the quality assessment of a 3 s block with a 1.4 average.

4.2.4 Normalization and Standardization

The next phase was, for each window, a normalization and standardization step. The normalization method used was the min-max normalization and results in a signal between $[0, 1]$. The expression for this normalization process can be seen in equation 4.3, where the $signal_{filt}$ represents the filtered signal and $signal_{norm}$ represents the normalized signal. A z-score standardization was used as the standardization step, which produces a signal with a mean of zero and a standard deviation of one. The expression for this step can be seen in 4.4, where $signal_{std}$ represents the standardized signal, μ the mean of the filtered signal, and σ the standard deviation of the filtered signal.

$$signal_{norm} = \frac{signal_{filt} - \min(signal_{filt})}{\max(signal_{filt}) - \min(signal_{filt})} \quad (4.3)$$

$$signal_{std} = \frac{signal_{filt} - \mu}{\sigma} \quad (4.4)$$

After this step, several features can be extracted. Depending on the feature, the raw, normalized or standardized signal was used. This can be seen in Section 4.3.

4.3 Feature Extraction

Several features were extracted from each window. Even though multiple signals were acquired, the features were only extracted from the SpO_2 sensor. This was done intentionally to minimize the use of external sensors. This independence simulates a wearable device and opens the possibility to, in the future, apply this algorithm to the sensor in specific. The features were extracted from the red and infrared channels separately, except the R ratio that uses information from both channels. Depending on the feature, the raw, normalized or standardized signal can be used. Table 4.2 shows which signal is used for each feature.

DC component: The DC component was determined by calculating the average of the unfiltered signal, represented by $signal^*$.

$$signal_{DC} = \frac{1}{N} \sum_{i=1}^N (signal_i^*) \quad (4.5)$$

Peak-to-peak component: The peak-to-peak (PTP) component was determined by subtracting the standardized signal maximum and minimum.

Table 4.2: Signal used for each extracted feature.

	Raw signal	Normalized signal	Standardized signal
DC component	x		
Peak-to-peak component			x
Mean		x	
Standard deviation		x	
Median			x
R ratio	x		x
S_{SQI}			x
K_{SQI}			x
P_{SQI}	x		x

$$signal_{PTP} = \max(signal_{std}) - \min(signal_{std}) \quad (4.6)$$

Mean and Standard Deviation: As the standardized signal has a mean value of zero and standard deviation of one, for both of this features the normalized signal was used.

Median: The signal median is the center value of signal, where the samples are organized in ascending order.

R ratio: This feature was calculated using the ratio presented in equation 2.4. This ratio is used to calculate the SpO_2 and combines the AC component (represented as $signal_{PTP}$) from both channels, after normalizing it by the DC component.

Skewness: The skewness measures the symmetry in a histogram, or the lack of it [77]. A histogram with zero skewness represents a perfectly symmetric distribution [106]. As represented in Figure 4.11, a histogram with a long right tail is positively skewed and a histogram with a long left tail is negatively skewed. Elgendi [76] showed that the skewness SQI (S_{SQI}) was not correlated with any other SQI tested. Furthermore, the S_{SQI} captured a unique perspective of the PPG signal. The expression used for computing the S_{SQI} is represented in equation 4.7, where N , μ and σ represent the number of samples, mean and standard deviation of the signal, respectively.

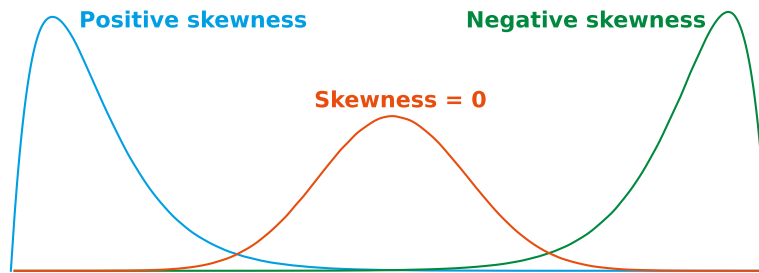


Figure 4.11: Histograms with positive, zero and negative skewness.

$$S_{SQI} = \frac{1}{N} \sum_{i=1}^N \left[signal_i - \frac{\mu}{\sigma} \right]^3 \quad (4.7)$$

Kurtosis: The kurtosis measures the peakness of a distribution [77]. A distribution is said to have positive kurtosis when it has a more pronounced peak and a heavy tail, when compared with a normal distribution, which has zero kurtosis. In contrast, a light tail and flatter distribution has a negative kurtosis. This is represented in Figure 4.12. Elgendi [76] also showed that the kurtosis SQI (K_{SQI}) was not correlated with any other SQI tested. The expression used for computing the K_{SQI} is represented in equation 4.8, where N , μ and σ represent the number of samples, mean and standard deviation of the signal, respectively.

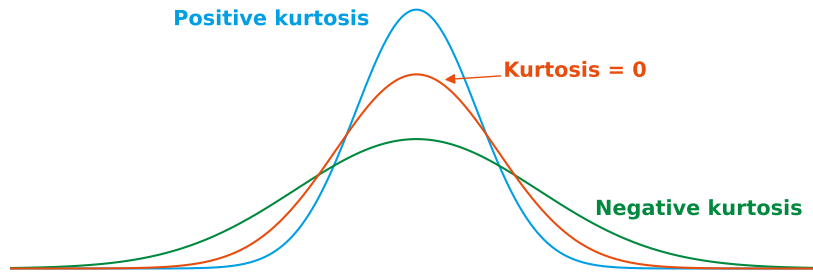


Figure 4.12: Histograms with positive, zero and negative kurtosis.

$$K_{SQI} = \frac{1}{N} \sum_{i=1}^N \left[signal_i - \frac{\mu}{\sigma} \right]^4 \quad (4.8)$$

Perfusion SQI: There are also several pulse oximetry manufactures that use the perfusion index (P_{SQI}) as the gold standard for assessing PPG quality [76]. This index can be expressed as the ratio of pulsatile blood (represented as the PTP component) to non-pulsatile blood (represented as the DC component), both calculated using the expressions 4.6 and 4.5.

$$P_{SQI} = \frac{signal_{PTP}}{signal_{DC}} \times 100 \quad (4.9)$$

4.3.1 Feature Selection

Not all features are equally important for a problem. An important step in ML is feature selection. It is an important and frequently used technique for dimension reduction by removing unnecessary features from data [44]. This approach can also provide a different understanding of the problem by ranking the different features.

A method based on Pearson correlation was used, calculated with Equation 4.10 where x and y represent the two variables and $\bar{x}$ and $\bar{y}$ the respective averages.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (4.10)$$

Correlation is a technique used to describe the relationship between two variables [107]. It is recommended to eliminate redundant or highly correlated features, as a dataset with correlated features increases computational complexity and can reduce the overall performance of the models [41, 43].

4.4 Machine Learning Models

In this work, for an automatic signal quality evaluation, two concepts were thought of: a classification approach and a regression approach. All models tested initially are presented in Table 2.1. After some initial tests, the RF model was chosen for further analysis. This decision was based on the fact that, for both classification and regression, this model had the best results.

As mentioned in Section 2.3.4, in clinical predictions the correct method for evaluate the ML models is a subject-wise cross-validation. Hence, in this work a 10-fold subject-wise cross-validation was used.

4.5 Physiological Metrics

Three clinical relevant metrics will be extracted from the PPG signal, the heart rate, respiratory rate and SpO₂. For the HR and RR estimation only one channel is needed. Hence, there can be more than one estimate for each window. In order to choose the final estimate three alternatives were developed:

- **Best (HR_{best} and RR_{best}):** This method picks the channel with the best quality. In case the signal quality is the same, the algorithm chooses the IR channel. This choice was done arbitrary.
- **Average (HR_{avg} and RR_{avg}):** Provided that the estimate was calculated for both channels, this method takes the average from both estimates. When only one channel has adequate quality, the algorithm outputs the only estimate possible.
- **Simulation (HR_{sim} and RR_{sim}):** This method simulates a real word scenario where the window quality from both channels is rounded to the nearest integer (simulating the classifiers developed). Then, it selects the channel with best quality. If both channels have the same quality, the algorithm averages the results.

4.5.1 Heart Rate Extraction

The heart rate is easily detected in the PPG signal as its periodicity is derived from the heartbeat [98]. The current gold standard method for HR estimation is the ECG [6],

therefore its results were used as the reference HR. This prediction was done using the windows previously used for quality assessment, i.e., three second windows with two seconds of overlap.

Reference HR (HR_{ref}): For estimation the HR using an ECG, typically the R-peaks are extracted from the signal. This was done using the BioSPPy Python package [108], which uses an algorithm developed by Christov [109]. Using the average difference of the peaks present in the window, the HR_{ref} was calculated.

Estimated HR (HR_{estm}): Firstly, a stricter filter was applied to the PPG signal. A 3rd order Butterworth band-pass filter was used. Several cutoff frequencies were tested, with the lower frequency being 0.5 Hz and the high frequency varying between 3 and 5 Hz. As in the quality assessment, the filter was applied forward and backward to the signal. Since the HR estimation does not need the signal features, only the main periodicity is essential, this filter could be applied to remove, in more depth, possible noise and false peaks.

After the new signal was obtained, the PPG peaks were determined using the BioSPPy Python package [108], which uses an algorithm developed by Elgendi et al. [110]. Using the average difference from the peaks present in the window, the HR was calculated for both the red (HR_{estm}^{red}) and infrared (HR_{estm}^{ir}) channels.

4.5.2 Respiratory Rate Extraction

In addition to HR, respiratory rate is another important parameter that can be estimated from the PPG signal. Multiple articles use capnometry as the reference method for RR estimation [69, 70, 92]. In this thesis this was not done and instead a RIP band was used, as shown in Figure 4.2. Due to the algorithms used and the acceptable RR frequencies, a larger window was needed. Hence, a 32 s sliding window with 31 s overlap was used.

Reference RR (RR_{ref}): Firstly, a 3rd order Butterworth band-pass filter with cutoff frequencies of 0.1 Hz and 2 Hz was applied forward and backward to the RIP signal. Then an adapted algorithm from the Biosppy package [108] was used to calculate RR_{ref} . The algorithm finds the zero-crossings of the standardized signal, ignoring the crossings that have a higher frequency than 1.3 Hz. As for every breath there is two zero-crossings, for every two crosses one is discarded, this can be seen in Figure 4.13. With the remaining points, the RR_{ref} is calculated based on the average difference between two consecutive breaths.

Estimated RR (RR_{estm}): The approach used to calculate the RR_{estm} was similar to the one used by Karlen et al. [69]. The algorithm was applied for both channels. Firstly, a digital filter was applied to the PPG signal. Since there are several approaches for filtering the PPG signal used in the different papers, such as low-pass filters [92], high-pass [69] and band-pass [70]. Along with the original filter, a high-pass filter with a cut-off frequency of 0.05 Hz and two band-pass filters with cut-off frequencies of 0.05 Hz and 5 Hz and 0.08 Hz and 3Hz were tested. Every filter was a 3rd order Butterworth

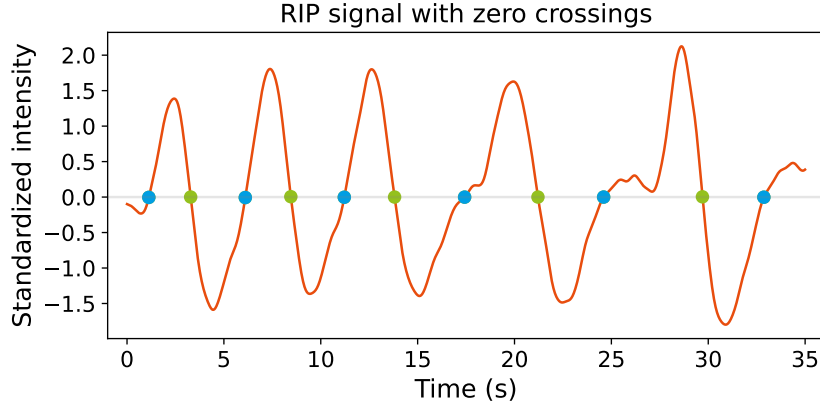


Figure 4.13: RIP signal with the zero-crossings highlighted. The crosses used for calculating the RR_{ref} represented in blue and the crosses ignored by the algorithm represented in green.

filter applied forward and backward to the signal. Then, the peaks and valleys of the PPG signal were found. This was done using the BioSPPy Python package [108], which uses an algorithm developed by Elgendi et al. [110]. Then, the three RIV mentioned in Section 3.2.4, RIAV, RIIV and RIFV, were obtained. These variations are then processed individually; this process can be seen in Figure 4.14.

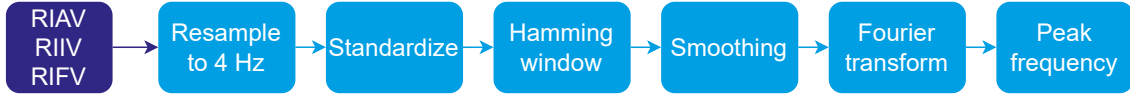


Figure 4.14: RIAV, RIIV and RIFV processing for RR estimation.

The variations are resampled to 4 Hz using linear interpolation, standardized, multiplied by a Hamming window and smoothed. The peak frequency in the expected respiratory frequency range (0.067-1.08 Hz) [69] is then calculated using a FFT. RR_{estm}^{red} and RR_{estm}^{ir} are the average of the accepted peak frequency of the three variations, for each channel. An example of a 10s window with the respective RIV and FFTs is illustrated in Figure 4.15. A 10s window is shown, instead of a 32s window, for a easier comprehension.

4.5.3 SpO_2 Extraction

Another important metric that can be extracted from a PPG sensor is the SpO_2 . As mentioned in Section 1.1, most of the wearable devices do not have medical certification, thus they cannot be used in clinical environments.

Reference SpO_2 ($SpO_{2\ ref}$): A commercially available pulse oximeter, Contec CMS50D+, was used as the reference method. The Contec CMS50D+ pulse oximeter can be integrated with the OpenSignals (r)evolution software. However, the signal is corrupted by high

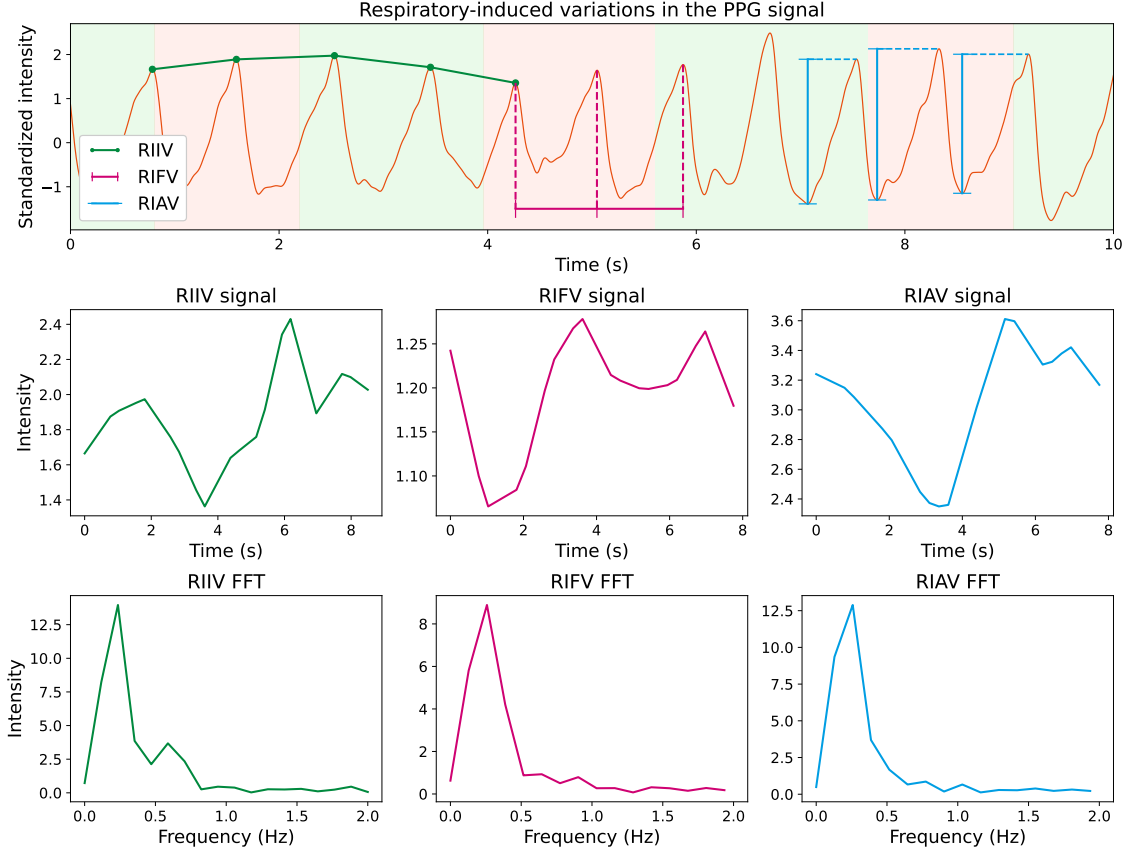


Figure 4.15: PPG signal with the respiratory-induced variations, and respective FFTs. On top it is represented an example of 10s window of the PPG signal with the respiratory-induced variations represented. The second and third line are the three RIV and their respective fast Fourier transform.

frequency noise. Therefore, a 5th order Butterworth low-pass filter with cutoff frequency of 1 Hz was applied and each value is rounded to the nearest integer, as this was the value presented in the device. As the $\text{SpO}_2_{\text{estm}}$ will be estimated pulse-by-pulse, the $\text{SpO}_2_{\text{ref}}$ is assumed to be the SpO_2 value at the pulse peak.

Estimated SpO_2 ($\text{SpO}_2_{\text{estm}}$): As mentioned in Section 2.2, the SpO_2 relates to the PPG signal via the ratio of ratios, R ratio, with its expression presented in Equation 2.4. This ratio uses the PPG signal AC and DC components from both channels, red and infrared, and it is calculated for every pulse. Afterwards, the R ratios in a determined window are averaged, along with the $\text{SpO}_2_{\text{ref}}$ for that window. Several window sizes were tested, in particular, 3, 5, 6, 9 and 10 s. An alternative R ratio calculation was also tested [111], with the expression presented in Equation 4.11, where λ_1 and λ_2 refers to both wavelengths used, in this work red and infrared, respectively.

$$R = \frac{\log\left(\frac{DC_{\lambda_1} + AC_{\lambda_1}}{DC_{\lambda_1}}\right)}{\log\left(\frac{DC_{\lambda_2} + AC_{\lambda_2}}{DC_{\lambda_2}}\right)} \quad (4.11)$$

The peaks and valleys were determined in the filtered signal using BioSPPy Python package [108], which uses an algorithm developed by Elgendi et al. [110]. Then, the AC component is assumed to be the difference between the values from the peak and the valley. The DC component is assumed to be the average value from the peak and valley, of the raw signal.

The R ratio and the $SpO_{2\text{ref}}$ are then used in order to produce a $SpO_{2\text{estm}}$. A Ridge regression was used. This is a linear regression method on which a penalty (P) is added to the Sum of Squared Errors (SSE), which is the parameter to minimize [45], as shown in Equation 4.12, where y_i and p_i represent the true and predicted values, respectively.

$$\text{minimize } (SSE + P) = \text{minimize } \left(\sum_{i=0}^n (y_i - p_i)^2 + P \right) \quad (4.12)$$

The penalty is controlled by a tuning parameter λ , where $\lambda \in [0, \infty[$. When $\lambda = 0$, there is no penalty applied and the objective function is a special case named Ordinary Least Squares (OLS) regression. The λ tested varied between 0 and 1, and 1 and 10 in a logarithmic scale.

Since the respiration was only suspended during a limited time, there is a limited amount of data with low values of SpO_2 , i.e., values lower than 90% SpO_2 . Therefore, a data balancing step was performed before the application of the linear regression. A random undersampling method was used, similar to one mentioned in Section 2.3.6. In this case, the data with SpO_2 values greater than 90% were undersampled.

4.5.4 Metrics Evaluation

As with regression problems, mentioned in Section 2.3.4.2, the values obtained from two different methods always implies some degree of error. Therefore to asses the performance of the developed algorithms some metrics from Section 2.3.4.2 will be used, along with the mean absolute percent error (MAPE) and the Bland-Altman Analysis [112].

Mean absolute percent error (MAPE): This metrics is a quantity used to measure the proximity the estimate is to the real value. This metrics is used by several studies [6] as well as recommended by the Consumer Technology Association [113].

$$MAPE = \frac{1}{N_{samples}} \sum_i^{N_{samples}} \left| \frac{y_i - p_i}{y_i} \right| \times 100\% \quad (4.13)$$

Where y_i represents the real output and p_i represents the predicted output.

Bland-Altman Analysis: This is the recommended method to determine agreement between medical instruments [6]. This method quantifies the agreement between two

measurements by calculating limits of agreement, calculated using the mean and the standard deviation of the differences between two measurements. The results are plotted, in which the Y axis represents the difference between the two measures and the X axis represents the average for these measures [114]. In the plot it is also represented the mean of the differences and an interval where 95% ($\mu \pm 1.96\sigma$) of the differences fall. If these intervals are within acceptable values, defined based on clinical necessity, the methods are considered to agree with each other. An example of a Bland-Altman plot for agreement between ECG and PPG can be seen in Figure 4.16.

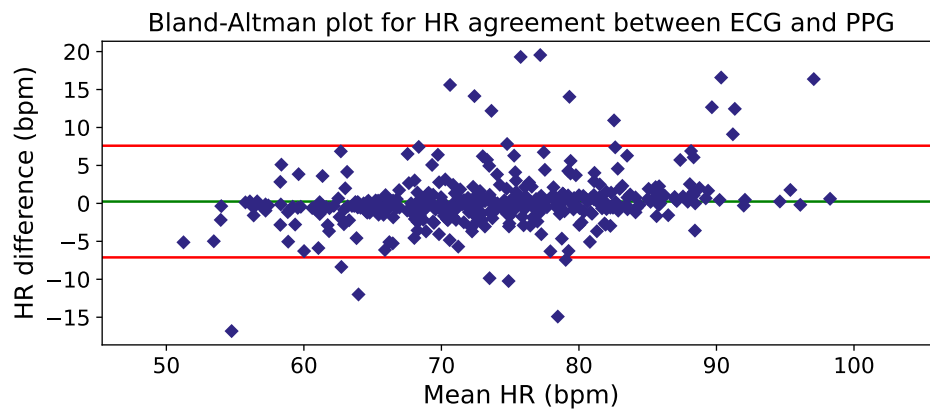


Figure 4.16: Bland-Altman plot for HR agreement between ECG and PPG. The line in green represents the average difference between the measurements (0.24 bpm) and in red the lines delimiting the 95% interval (± 7.36 bpm).

Results

This Chapter presents the results achieved in this thesis. Firstly, a summary of the data acquired during this work is provided. Subsequently, the results achieved in the manual classification, the regression and classification models performance and the features importance are also displayed. Finally, the heart rate, respiratory rate and SpO₂ extraction results are also presented.

5.1 Data Acquisition

5.1.1 Acquisition Protocol for Quality Assessment

For the procedure described in Section 4.1.1, five individuals (4 men), aged between 20 and 23 years old (mean = 21.4, standard deviation = 1.0) were recruited. All participants were right-handed and all subjects, except one, wore the devices on the left wrist. There were no subjects with any medical conditions relevant.

The mean total acquisition time, i.e., the time between turning on and off the recording software, was 22 minutes and 34 seconds (with a standard deviation of 3 minutes and 37 seconds). The action mean total time, i.e. the sum of moments performing the actions represented in Figure 4.3, was 17 minutes and 11 seconds (with a standard deviation of 3 minutes and 33 seconds).

5.1.2 Acquisition Protocol for SpO₂ Extraction

The protocol described in Section 4.1.2 was designed with tasks less prone to motion artifacts and two breathing exercises, in order to cause a decrease in SpO₂ values. An example of the SpO₂ evolution can be seen in Figure 5.1. Several acquisitions were made, all to the same individual, and only the signals acquired in the breathing exercises were used, totaling, approximately, 870 s of usable data. As can be seen in Figure 5.1, most of the time the subject has an elevated value of SpO₂, this results in 89% of all data having SpO₂ > 90%, which led to the development of a balancing step prior to the linear regression.

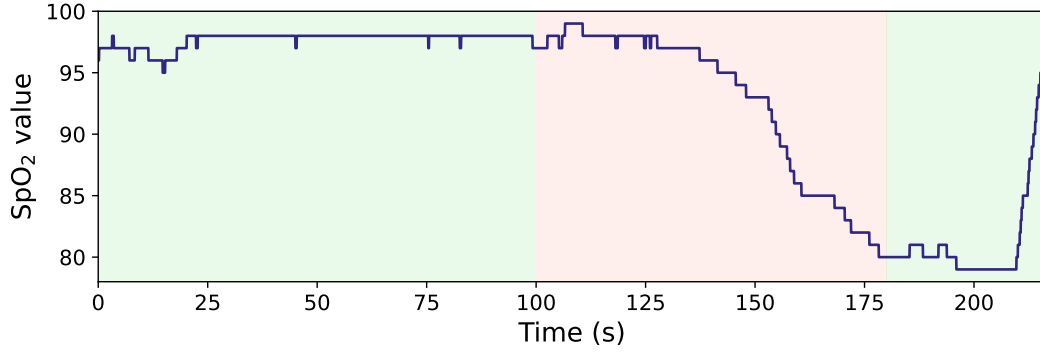


Figure 5.1: SpO_2 evolution during a breathing exercise, where the green and red backgrounds represent the breathing and breath hold phases, respectively. The breath was hold during approximately 80 s, lowering the SpO_2 value down to a minimum of 79%.

5.2 Quality Assessment

As mentioned in Section 4.2.3, all data was manually evaluated regarding its quality. This evaluation was done by two different annotators and a inter-annotator agreement measure was defined. This was evaluated in a 2 min portion of data. A Cohen's kappa of 0.813 and of 0.971 was obtained for the red and infrared channels, respectively.

Depending on the interpretation, there is a slight variation on the prevalence of each quality. These results can be seen in Table 5.1. It can be noted that with the 'Strict Average' quality a large portion of the windows are classified as low-quality. This is a result of the threshold imposed. Therefore, this interpretation will not be used throughout the development of the different models.

Table 5.1: Results of the manual quality evaluation indicating the prevalence of the different interpretations.

		Samples	'Average' quality	'Strict Average' quality	'Mode' quality
Red Signal	Class '0'	74.00 %	73.51 %	93.75 %	74.51 %
	Class '1'	17.09 %	18.57 %	5.05 %	16.70 %
	Class '2'	8.91 %	7.92 %	1.20 %	8.79 %
Infrared Signal	Class '0'	61.65 %	60.36 %	92.45 %	61.64 %
	Class '1'	19.18 %	21.70 %	5.60 %	19.20 %
	Class '2'	19.17 %	19.94 %	19.95 %	19.16 %

In Table 5.2 it is possible to examine the samples distribution of each quality for all subjects. It is possible to note that Subject 4 has very low quality data, with a large portion of the data evaluated as Class '0' and little or no data evaluated as Class '2'. Subject 3 is in a similar situation. During the training of the classifiers, when balancing the data, this would result in the classifiers being trained/tested with very little or, in some cases, with

no samples of good quality. For this reason, both of these subjects were excluded from the ML development.

Table 5.2: Results of the manual quality evaluation, by subject.

		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Red Signal	Class '0'	40.91 %	74.44 %	75.98 %	96.00 %	70.48 %
	Class '1'	14.51 %	23.84 %	22.99 %	4.00 %	17.30 %
	Class '2'	44.58 %	1.72 %	1.03 %	0.00 %	12.21 %
Infrared Signal	Class '0'	29.96 %	59.58 %	63.10 %	87.21 %	57.17 %
	Class '1'	2.73 %	31.51 %	33.38 %	12.15 %	6.22 %
	Class '2'	67.31 %	8.91 %	3.52 %	0.64 %	36.61 %

5.2.1 Regression Models

This section presents the results for the regression models developed. Only the '*Average*' quality was used for developing this models. This was done since the '*Strict Average*' quality will not be used in any development, as mentioned in Section 5.2, and the '*Mode*' quality only outputs discrete qualities, therefore, it is more appropriate to use a classification approach.

Firstly, different models, which can be seen in Table 2.1, were tested to select the better performing out-of-the-box, the results can be seen in Table 5.3. These models were tested with their default parameters [48]. It is possible to note that the Random Forest and Gradient Boosting models have very similar results, hence, both models could be further analyzed with more detail. The RF model was chosen, since a similar model was used in the classification problems.

Table 5.3: Performance results for the different regression models.

Regressor	Red channel			IR channel		
	MAE	RMSE	R ²	MAE	RMSE	R ²
Ridge	0.40±0.07	0.72±1.09	-5.34±28.61	0.48±0.06	0.87±1.10	-5.47±25.99
Lasso	0.55±0.07	0.61±0.08	-0.05±0.12	0.63±0.08	0.68±0.08	-0.09±0.07
SVR	0.57±0.13	0.65±0.16	-0.21±0.41	0.61±0.12	0.69±0.13	-0.09±0.17
KNNR	0.65±0.06	0.84±0.05	-1.03±0.41	0.73±0.07	0.88±0.05	-0.93±0.56
DTR	0.47±0.10	0.67±0.10	-0.27±0.32	0.53±0.06	0.71±0.06	-0.25±0.35
RFR	0.39±0.09	0.51±0.10	0.26±0.20	0.45±0.06	0.55±0.06	0.28±0.08
GBR	0.40±0.10	0.52±0.11	0.22±0.26	0.45±0.06	0.54±0.06	0.31±0.10

The Random Forest models were tested with different parameters, in order to find the optimal combination. The parameters chosen were the maximum depth and the number of trees generated. The maximum depths tested were from 1 to 10 with unitary jumps and from 10 to 100 with jumps of 10. The number of trees generated tested were between

1 and 100 in a logarithmic scale and 200, 500 and 800. As in this work a 10-fold subject-wise cross-validation was used and there were three subjects used for developing the ML models, each combination was tested 30 times.

Figure 5.2 shows the regression results for the red and infrared channels in function of the number of trees generated and the maximum depth. Even though there were models tested with 800 trees and 100 maximum depth, in Figure 5.2 the limit is 100 trees and 50 maximum depth, respectively. This was done to simplify the graphics, since there was no change in the behavior.

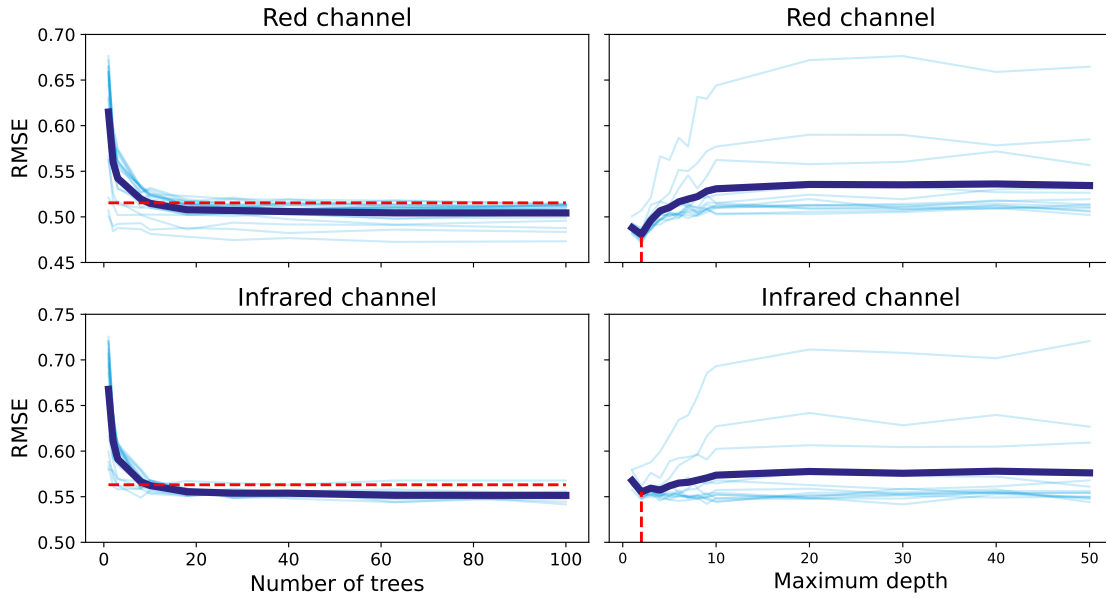


Figure 5.2: RMSE results for the red (top plots) and IR channels (bottom plots) in function of the number of trees generated (left plots) and the maximum depth (right plots). The lighter blue lines represent the performance for the different combinations. The dark blue line represents the average for every number of trees generated or for every maximum depth, for the left and right graphics respectively. On the left graphics the red line represents 90% of the RMSE improvement, represented in 0.515 and 0.563 for the red and IR channel, respectively. On the right graphics the red line represents the minimum RMSE, represented in the maximum depth 2 for both channels.

Table 5.4 shows the results for the best combinations of red and infrared regression models. Since in a wearable device performance is a limiting factor, the best models for each channel and quality interpretation shown are limited to 100 generated trees and a maximum depth of 50. These values were chosen given the fact that Figure 5.2 shows no significant improvements after these values.

Figure 5.3 shows the real and predicted red quality for a 10-fold subject-wise cross-validation for the model presented in Table 5.4.

Table 5.4: Performance results for the best combinations for the regression models.

	Red channel	IR channel
Trees	100	100
Depth	2	30
MAE	0.37 ± 0.05	0.44 ± 0.06
RMSE	0.47 ± 0.06	0.54 ± 0.06
R^2	0.37 ± 0.10	0.31 ± 0.06

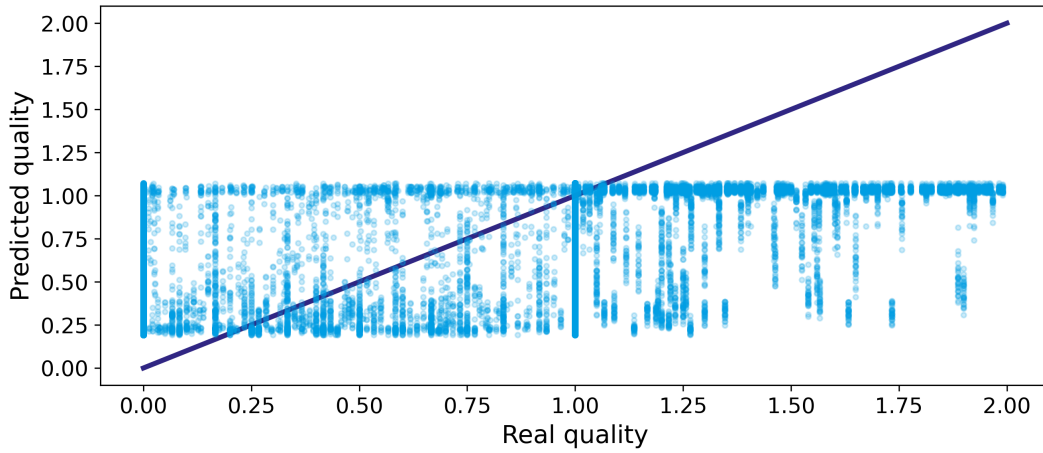


Figure 5.3: Real and predicted red quality for the best red channel regressor. The dark blue line represents a perfect regression and the lighter blue dots represent the evaluations made by the Random Forest regressor.

5.2.2 Classification Models

This section presents the results for the classification models developed. Firstly, different models, which can be seen in Table 2.1, were tested and their results can be seen in Table 5.5. These models were tested with their default parameters [48] and with the 'Average' quality. It is possible to note that the RF model had the best results and thus this model was chosen for further optimization. Similar to the regression models mentioned in section 5.2.1, the Random Forest models were tested with different parameters. These parameters were the same used in the regression models.

As mentioned in Section 4.3, the features were extracted from the red and infrared channels separately, except the R ratio that uses information from both channels. Therefore there were two scenarios for which features to use: (1) use features from both channels to train both classifiers; (2) train each channel classifier with features only from the respective channel. In the second possibility the R ratio was not used, since it always uses information from both channels.

Table 5.5: Performance results for the different classification models.

Classifier	Red channel			IR channel		
	Accuracy	Precision	F-score	Accuracy	Precision	F-score
KNNC	0.31±0.06	0.18±0.08	0.21±0.06	0.29±0.08	0.22±0.14	0.21±0.08
DTC	0.50±0.07	0.54±0.09	0.48±0.08	0.52±0.07	0.55±0.05	0.50±0.07
RFC	0.60±0.07	0.68±0.07	0.59±0.08	0.61±0.06	0.66±0.05	0.60±0.06
SVC	0.32±0.07	0.33±0.15	0.25±0.06	0.30±0.06	0.24±0.14	0.19±0.03
GNB	0.50±0.05	0.53±0.07	0.47±0.07	0.48±0.07	0.51±0.09	0.45±0.07
MLP	0.33±0.05	0.16±0.12	0.18±0.06	0.34±0.06	0.21±0.15	0.20±0.08

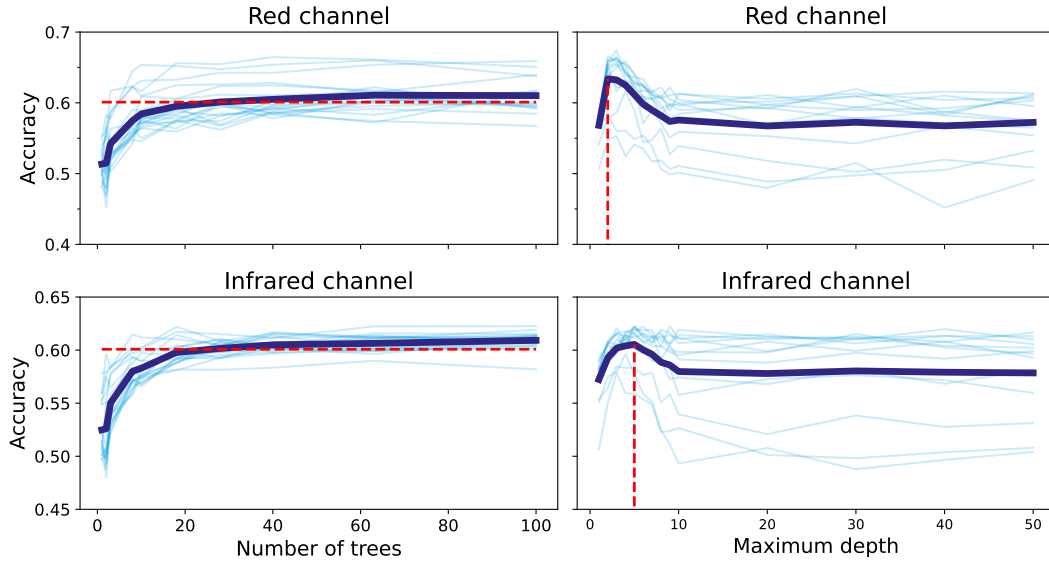
5.2.2.1 Features From Both Channels

Figure 5.4 shows the accuracy results for the red and infrared classifiers in function of the number of trees generated and the maximum depth, for the two quality interpretations used. It is possible to note that the different models, with different parameters combinations, have similar behaviors, despite having different values. Only the accuracy values for the average are shown for simplicity, but it is possible to note that there are models that perform noticeably above the average.

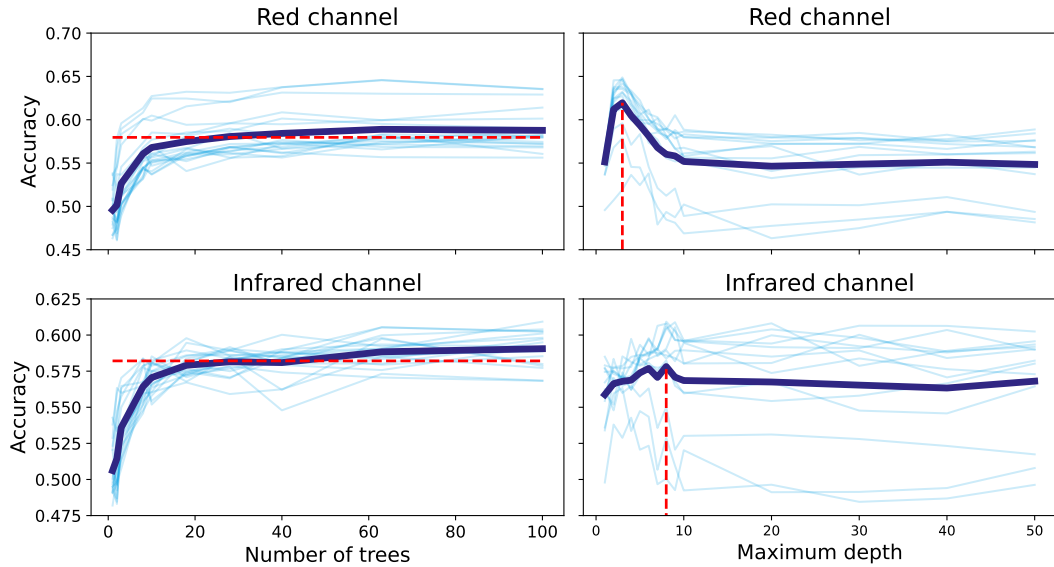
Table 5.6 shows the performance results for the best model combinations developed using the features from both channels and Figure 5.5 shows the confusion matrices for these models. Since in a wearable device performance is a limiting factor, the best models for each channel and quality interpretation shown are limited to 100 generated trees and a maximum depth of 50. These values were chosen given the fact that Figure 5.4 shows no significant improvements after these values.

Table 5.6: Performance results for the best combinations for the classification models, using features from both channels.

	Red channel		IR channel	
	'Average' quality	'Mode' quality	'Average' quality	'Mode' quality
Trees	40	63	100	100
Depth	3	3	5	8
Accuracy	0.67±0.05	0.65±0.05	0.62±0.04	0.61±0.06
Precision	0.70±0.04	0.69±0.04	0.65±0.05	0.65±0.08
Recall	0.67±0.05	0.65±0.05	0.62±0.04	0.61±0.06
F-score	0.66±0.05	0.64±0.06	0.62±0.04	0.59±0.06
AUC-ROC	0.83±0.02	0.81±0.02	0.80±0.03	0.78±0.05



(a) 'Average' quality.



(b) 'Mode' quality.

Figure 5.4: Accuracy results for the red (top plots) and IR (bottom plots) channels in function of the number of trees generated (left plots) and the maximum depth (right plots). Models developed using all features. The lighter blue lines represent the performance for the different combinations. The dark blue line represents the average for every number of trees generated or for every maximum depth, for the left and right graphics respectively. On the left graphics the red line represents 90% of the accuracy improvement, represented in 0.60 for the 'Average' quality and 0.58 for the 'Mode' quality. On the right graphics the red line represents the highest accuracy, represented in the maximum depth 3 and 5 for red and infrared channel in both quality interpretations.

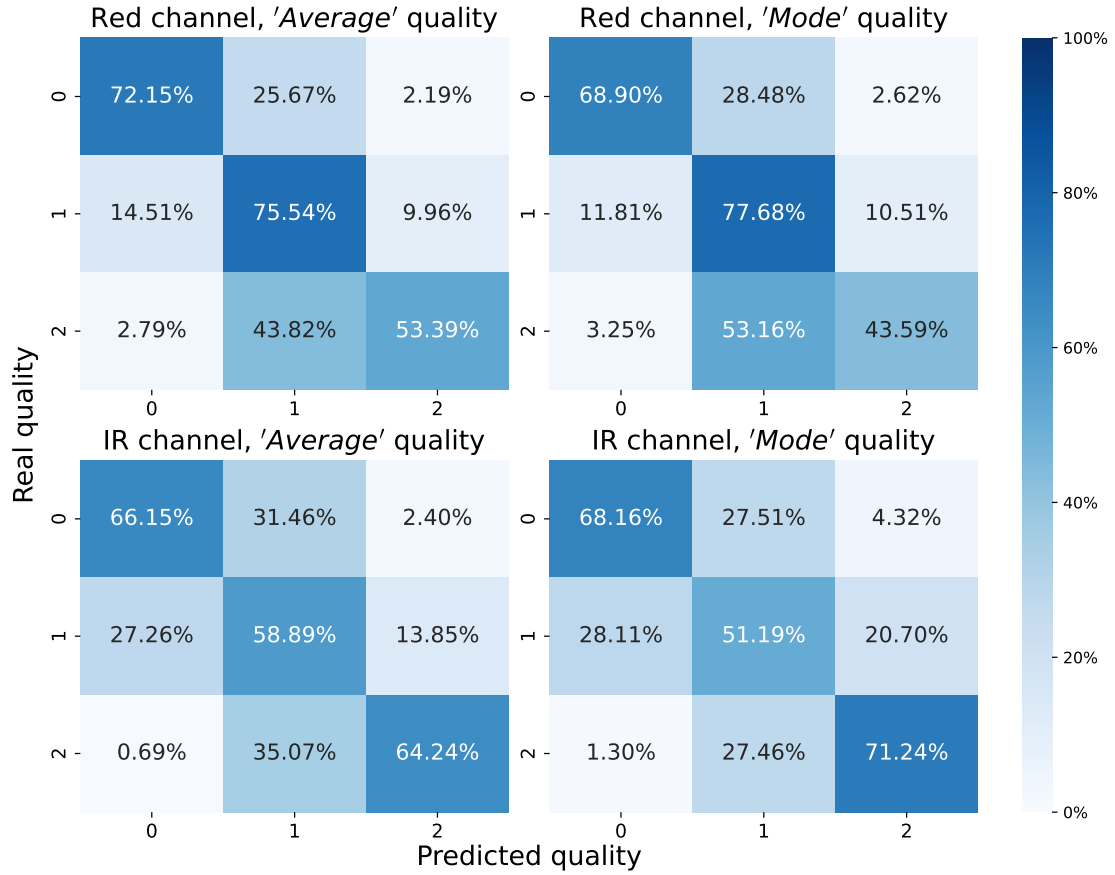
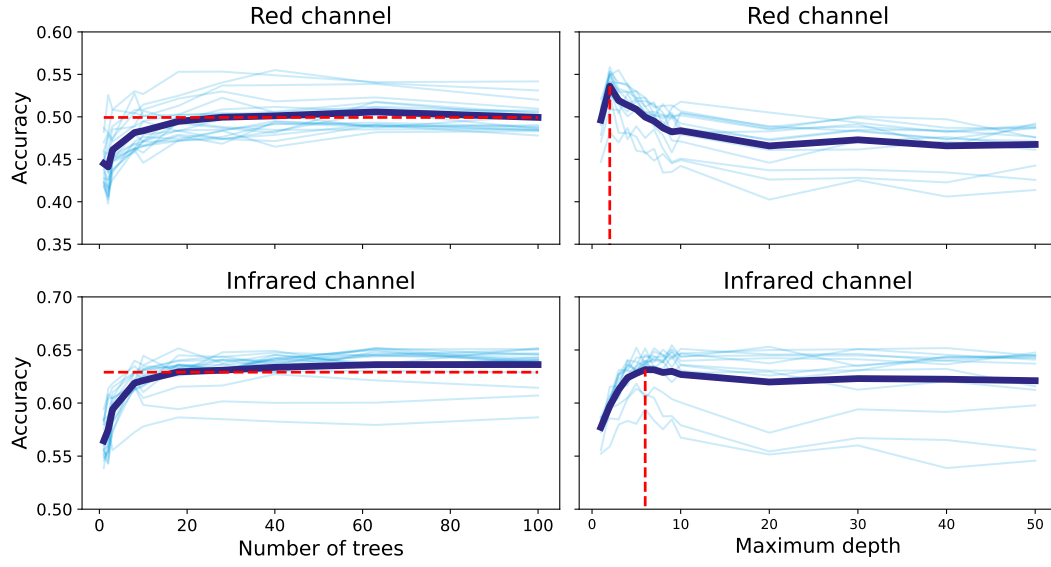


Figure 5.5: Confusion matrices for the best classification models for the red and IR channels, using the two different quality approaches and features from both channels.

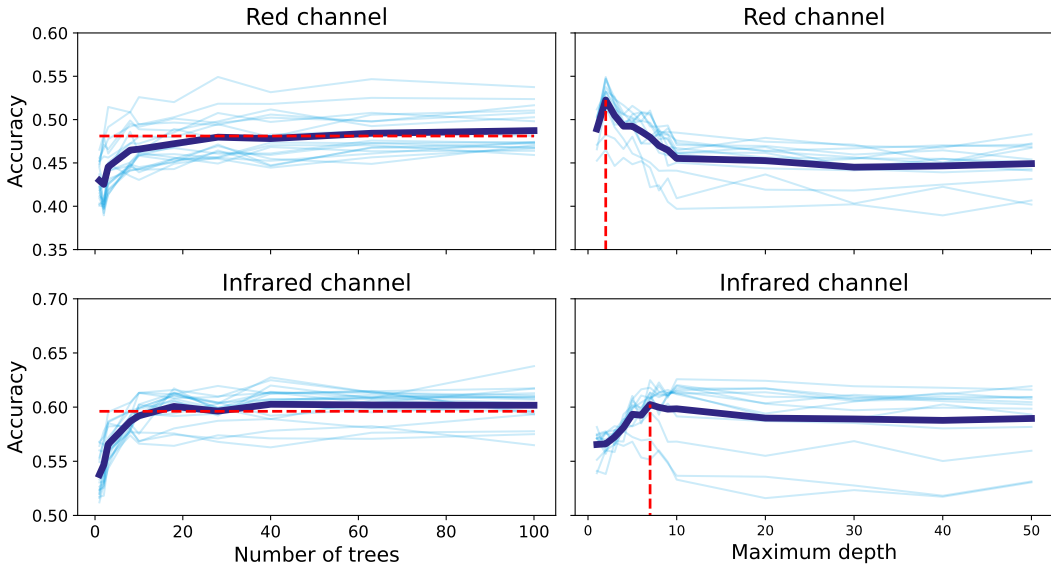
5.2.2.2 Features From Individual Channels

Figure 5.6 shows the accuracy results for the red and infrared classifiers in function of the number of trees generated and the maximum depth, for both interpretations of the quality used. It is possible to note that the different models, with different parameters combinations, have similar behaviors, despite having different values. Only the accuracy values for the average are shown for simplicity, but it is possible to note that there are models that perform noticeably above the average.

Table 5.7 shows the performance results for the best model combination developed using the features from individual channels and Figure 5.7 shows the confusion matrices for these models. Since in a wearable device performance is a limiting factor, the best models for each channel and quality interpretation shown are limited to 100 generated trees and a maximum depth of 50. These values were chosen given the fact that Figure 5.6 shows no significant improvements after these amounts.



(a) 'Average' quality.



(b) 'Mode' quality.

Figure 5.6: Accuracy results for the red (top plots) and IR channels (bottom plots) in function of the number of trees generated (left plots) and the maximum depth (right plots). Models developed using only features from the respective individual channel. The lighter blue lines represent the performance for the different combinations. The dark blue line represents the average for every number of trees generated or for every maximum depth, for the left and right graphics respectively. On the left graphics the red line represents 90% of the accuracy improvement, represented in 0.50 and 0.63 for the 'Average' quality and 0.48 and 0.60 for the 'Mode' quality, for the red and IR channels respectively. On the right graphics the red line represents the highest accuracy, represented in the maximum depth 2 and 6 for the 'Average' quality and 2 and 7 for the 'Mode' quality, for the red and IR channels respectively

Table 5.7: Performance results for the best combinations for the classification models, using features from individual channels.

	Red channel		IR channel	
	'Average' quality	'Mode' quality	'Average' quality	'Mode' quality
Trees	40	28	63	40
Depth	3	2	40	7
Accuracy	0.56±0.05	0.55±0.03	0.65±0.03	0.63±0.06
Precision	0.60±0.07	0.56±0.04	0.66±0.04	0.63±0.07
Recall	0.56±0.05	0.55±0.03	0.65±0.03	0.63±0.06
F-score	0.55±0.05	0.54±0.03	0.65±0.04	0.61±0.06
AUC-ROC	0.73±0.06	0.70±0.02	0.81±0.03	0.79±0.04

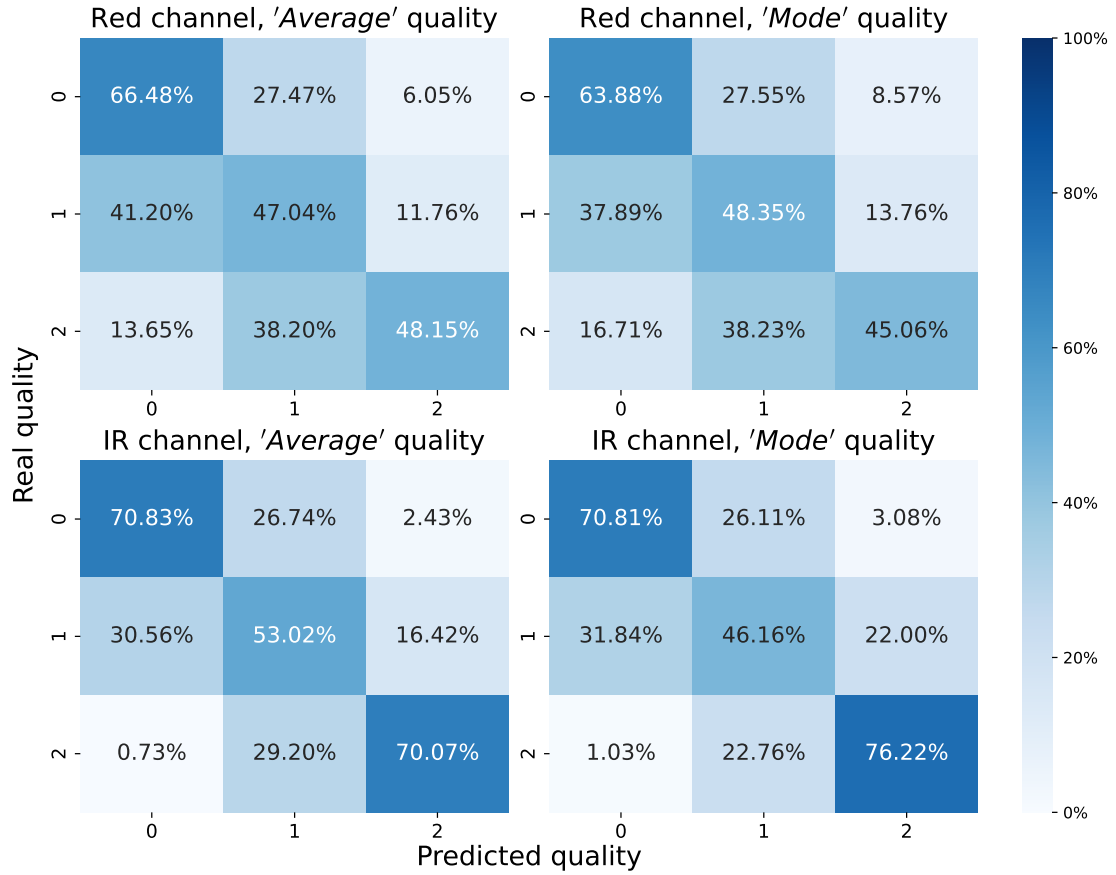


Figure 5.7: Confusion matrices for the best classification models for the red and IR channels, using the two different quality approaches and features from individual channels.

5.2.2.3 Double Classifiers

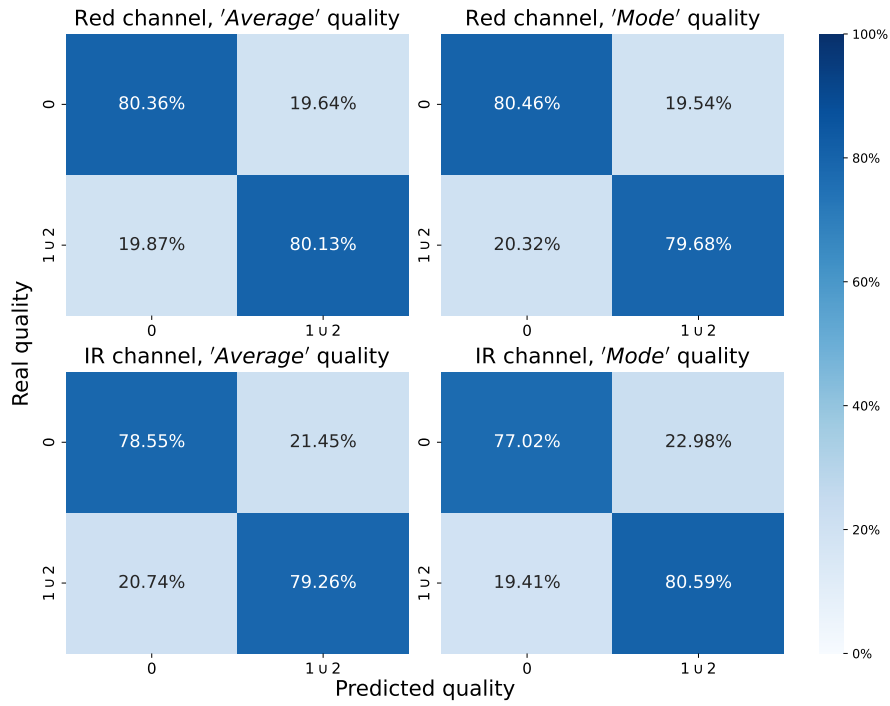
Another alternative studied was the classification in two steps, i.e., one classifier to separate the data between classes '0' and '1 $\cup$ 2' (where this second class indicates that the window can have either quality '1' or '2'), and second classifier to evaluate between the classes '1' and '2'. This was designed since the multi-class classifiers showed some difficulty dividing between the classes '1' and '2', thus emphasizing this partition. The results for the best combinations can be seen in Table 5.8 and 5.9, respectively. Figure 5.8 shows the confusion matrices for these models.

Table 5.8: Performance results for the best combinations for the classification models differentiating between class '0' and '1 $\cup$ 2', using features from both channels.

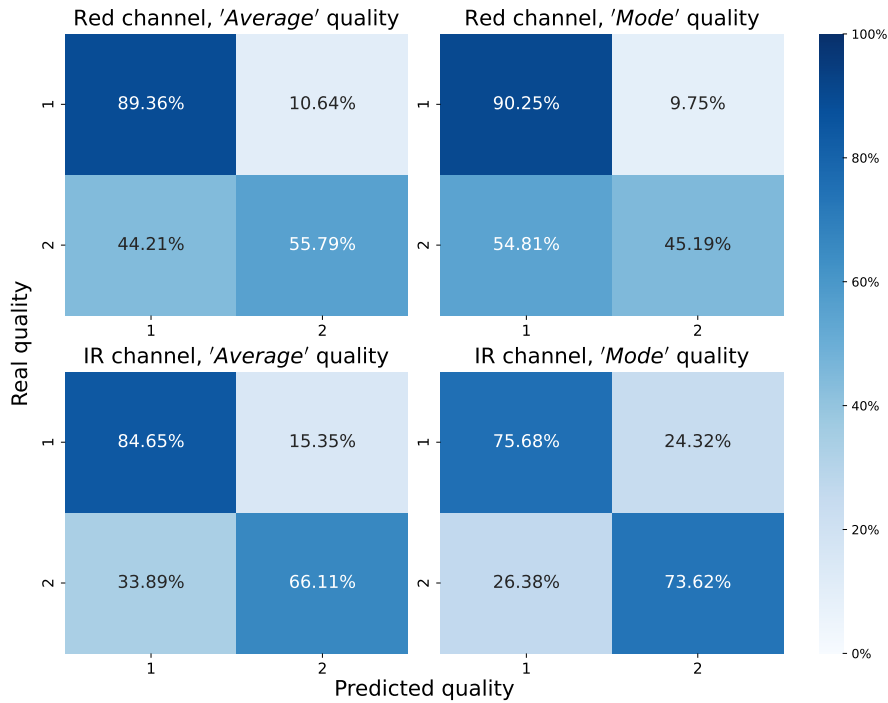
	Red channel		IR channel	
	'Average' quality	'Mode' quality	'Average' quality	'Mode' quality
Trees	63	100	63	100
Depth	3	5	7	5
Accuracy	0.80 $\pm$ 0.04	0.80 $\pm$ 0.04	0.81 $\pm$ 0.07	0.80 $\pm$ 0.04
Precision	0.80 $\pm$ 0.04	0.81 $\pm$ 0.03	0.79 $\pm$ 0.02	0.81 $\pm$ 0.03
Recall	0.80 $\pm$ 0.12	0.79 $\pm$ 0.12	0.83 $\pm$ 0.16	0.79 $\pm$ 0.12
F-score	0.80 $\pm$ 0.06	0.79 $\pm$ 0.06	0.80 $\pm$ 0.09	0.79 $\pm$ 0.06
AUC-ROC	0.87 $\pm$ 0.06	0.86 $\pm$ 0.06	0.87 $\pm$ 0.08	0.86 $\pm$ 0.06

Table 5.9: Performance results for the best combinations for the classification models differentiating between class '1' and '2', using features from both channels.

	Red channel		IR channel	
	'Average' quality	'Mode' quality	'Average' quality	'Mode' quality
Trees	100	100	28	100
Depth	2	2	4	8
Accuracy	0.74 $\pm$ 0.09	0.72 $\pm$ 0.10	0.75 $\pm$ 0.08	0.81 $\pm$ 0.07
Precision	0.70 $\pm$ 0.10	0.68 $\pm$ 0.09	0.75 $\pm$ 0.14	0.78 $\pm$ 0.02
Recall	0.89 $\pm$ 0.07	0.89 $\pm$ 0.07	0.81 $\pm$ 0.10	0.84 $\pm$ 0.17
F-score	0.78 $\pm$ 0.06	0.77 $\pm$ 0.06	0.76 $\pm$ 0.06	0.80 $\pm$ 0.09
AUC-ROC	0.80 $\pm$ 0.07	0.77 $\pm$ 0.09	0.83 $\pm$ 0.07	0.87 $\pm$ 0.08



(a) Models for differentiating between class '0' and '1 ∪ 2'



(b) Models for differentiating between class '1' and '2'

Figure 5.8: Confusion matrices for the best classification models for the red and IR channels for the double classifiers.

5.2.2.4 Vote Classifiers

Another alternative studied was the use of multiple classifiers that vote on a final classification, as an ensemble learning method. The classifiers selection was done based on the ones that perform better than random selecting a class (accuracy > 33%) with the default parameters. This type of classifier was tested using the features from both channels and the two quality interpretations.

Three classifiers were chosen: (1) a Random Forest model with the best accuracy, represented in Table 5.6; (2) a Gaussian Naive Bayes classifier, with default parameters [48]; (3) a single decision tree with different parameters tested. These parameters were: (a) the function to measure the quality of a split, with the 'gini' and 'entropy' functions tested; (b) the maximum depth of a tree, with depths from 1 to 10 with unitary jumps and from 10 to 100 with jumps of 10; (c) the minimum number of samples required to split an internal node, with values from 2 to 9 tested.

The results for the best combinations can be seen in Table 5.10. The AUC-ROC was not calculated for this alternative. Figure 5.9 shows the confusion matrices for these models.

Table 5.10: Performance results for the best combinations of the vote classification models, using features from both channels.

	Red channel		IR channel	
	'Average' quality	'Mode' quality	'Average' quality	'Mode' quality
Split function	Gini	Gini	Entropy	Entropy
Depth	2	2	1	2
Min. samples	3	8	7	4
Accuracy	0.66±0.06	0.64±0.07	0.63±0.05	0.61±0.06
Precision	0.69±0.06	0.67±0.05	0.60±0.05	0.61±0.08
Recall	0.66±0.06	0.64±0.07	0.63±0.04	0.61±0.06
F-score	0.66±0.07	0.63±0.08	0.58±0.04	0.59±0.06

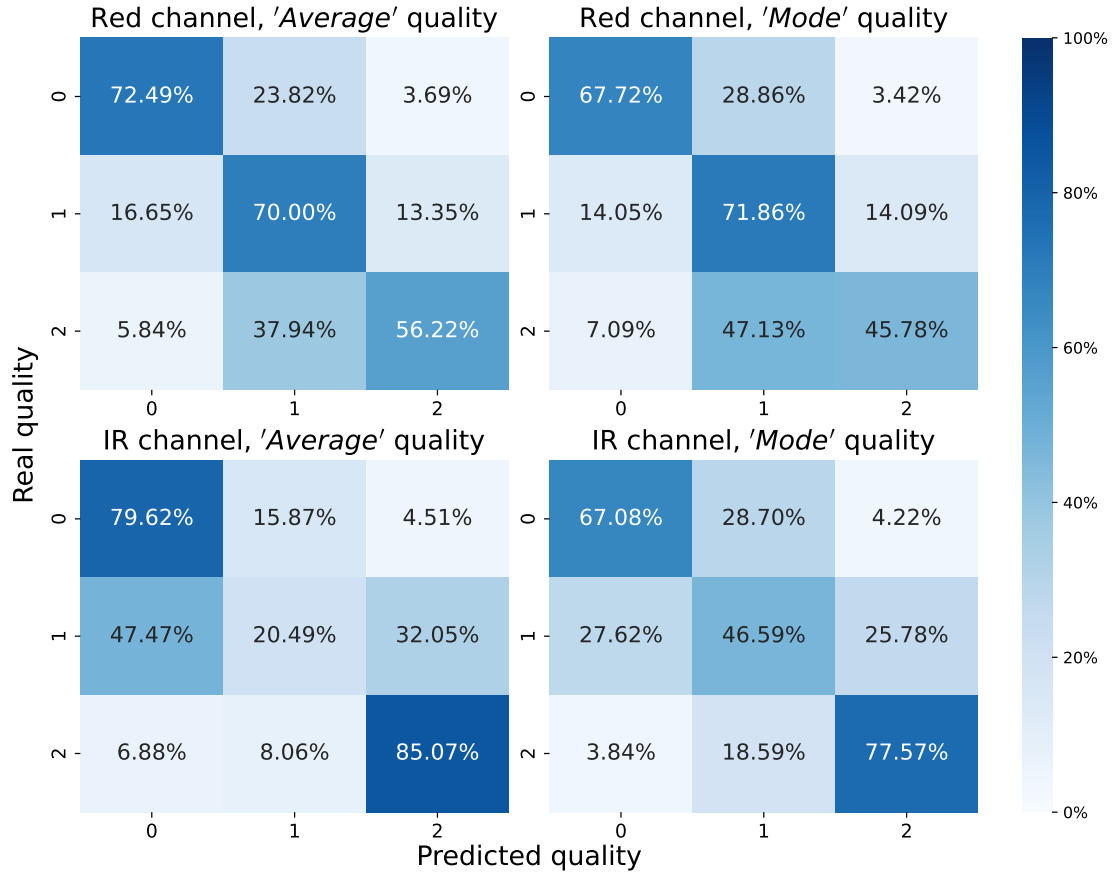


Figure 5.9: Confusion matrices for the best vote classification models for the red and IR channels, using the two different quality approaches and features from both channels.

5.2.3 Feature Importance

Pearson correlation was used for describing the relationship between the quality of the signal and their features. Figure 5.10 shows the correlations between all features and each quality.

Figure 5.11 represents the accuracy for the RF models trained with different number of features. The Random Forest models parameters chosen were defined in Table 5.6, and they were trained with a 10-fold subject-wise cross-validation. The order by which the features were selected is defined in Figure 5.10.

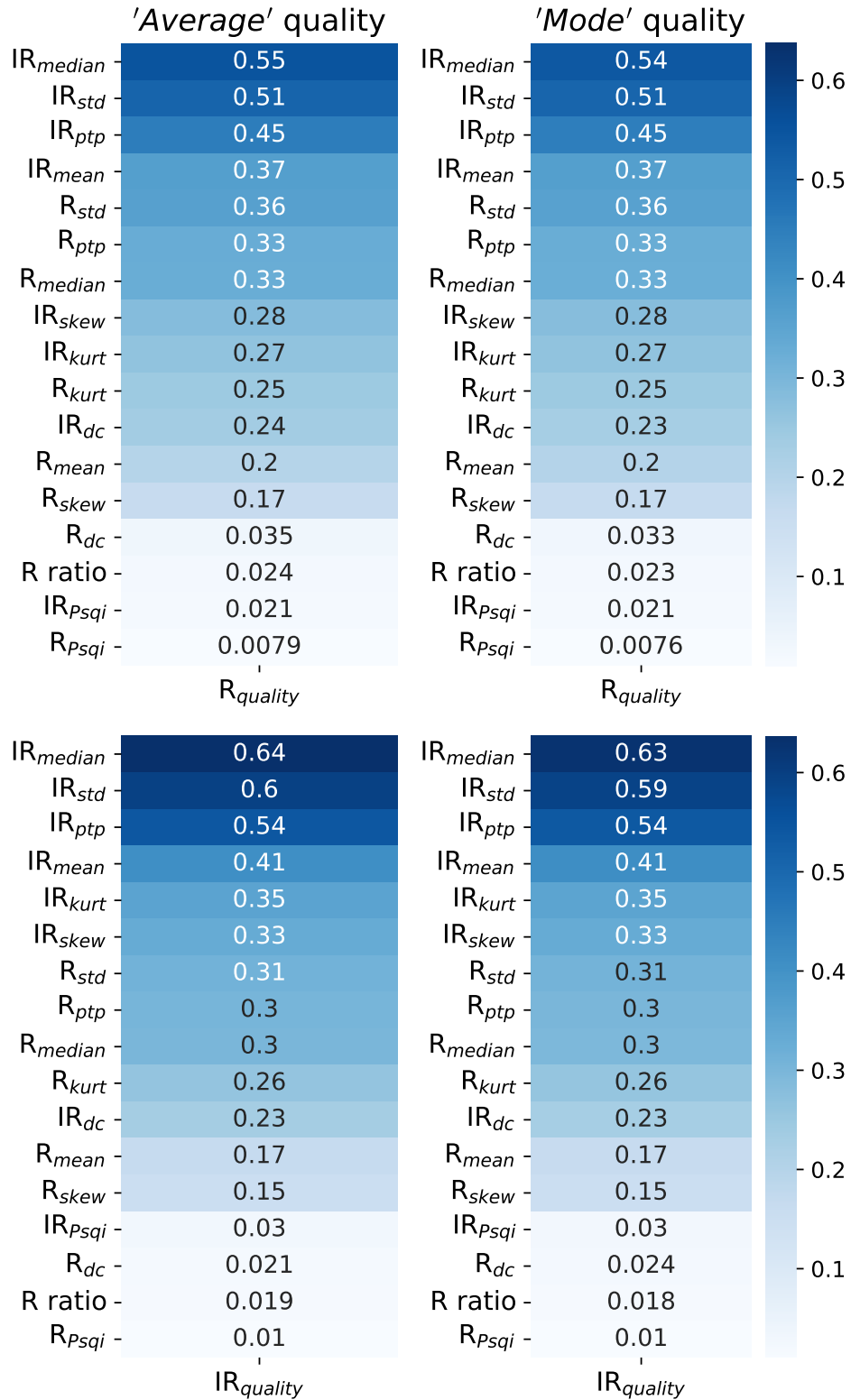
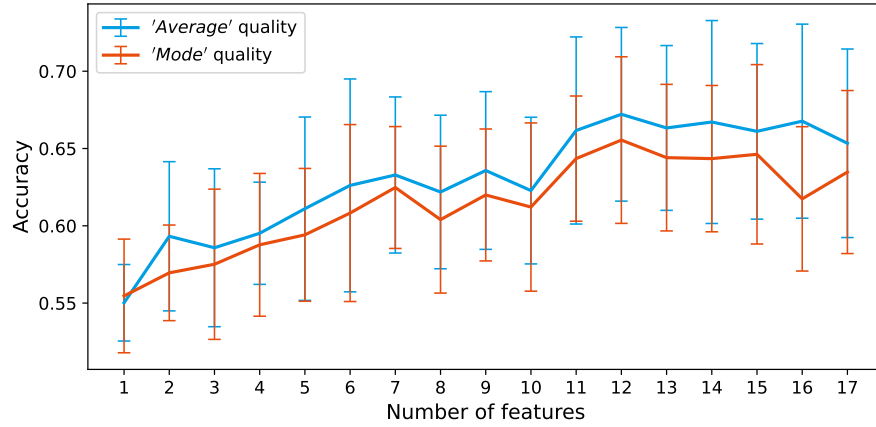
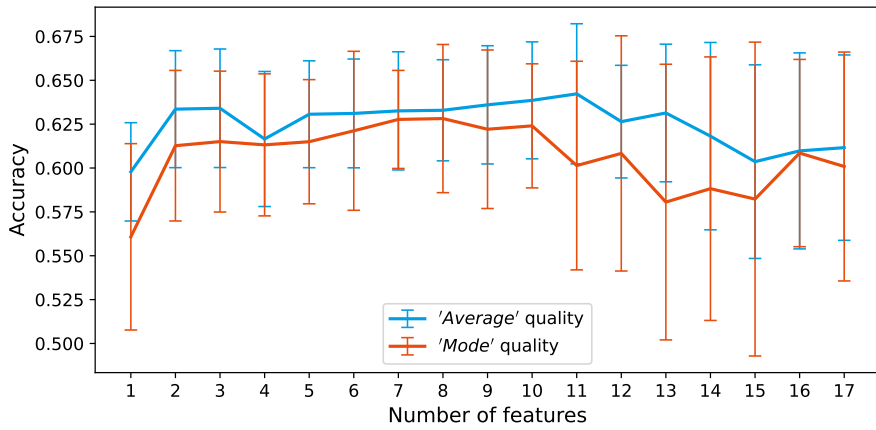


Figure 5.10: Correlation between all the features and the quality from the red and IR channels, for both quality interpretations.



(a) Red channel.



(b) Infrared channel.

Figure 5.11: Accuracy for models trained with different amounts of features, where the 'Average' and 'Mode' qualities are represented in blue and orange, respectively, and the error bars represent the standard deviation. The order by which the features were selected was based on the correlation between the feature and the signal quality.

5.3 Physiological Metrics

5.3.1 Heart Rate Extraction

This section presents the results from the HR extraction. The results presented in this section can use two types of data: (1) at least one channel has quality '2'; (2) at least one channel has quality '1' or better. Windows where both channels are classified as quality '0' were not used since it is assumed that the signal has no relevant information that can be extracted.

There were six different filters tested for the HR estimation, the original filter and 5 alternatives with different cutoff frequencies. Table 5.11 shows the MAE and MAPE results when utilizing windows including at least one channel with quality '1' or better.

Table 5.12 shows the results with data from windows including at least one channel with quality '2'.

Table 5.11: Results for the three approaches for HR estimation with at least one channel with quality '1' or better.

High cutoff frequency	MAE _{best} (bpm)	MAE _{avg} (bpm)	MAE _{sim} (bpm)	MAPE _{best}	MAPE _{avg}	MAPE _{sim}
3 Hz	4.44	5.18	4.88	5.28 %	6.25 %	5.84 %
4 Hz	4.23	4.95	4.67	5.02 %	5.96 %	5.59 %
5 Hz	4.04	4.82	4.53	4.79 %	5.81 %	5.43 %
6 Hz	4.09	4.86	4.57	4.88 %	5.90 %	5.51 %
7 Hz	4.05	4.85	4.55	4.82 %	5.88 %	5.48 %
Original	4.37	5.19	4.86	5.22 %	6.29 %	5.85 %

Table 5.12: Results for the three approaches for HR estimation with at least one channel with quality '2'.

High cutoff frequency	MAE _{best} (bpm)	MAE _{avg} (bpm)	MAE _{sim} (bpm)	MAPE _{best}	MAPE _{avg}	MAPE _{sim}
3 Hz	1.28	2.08	1.40	1.72 %	2.78 %	1.89 %
4 Hz	1.24	1.95	1.34	1.67 %	2.61 %	1.81 %
5 Hz	1.23	1.99	1.37	1.65 %	2.67 %	1.84 %
6 Hz	1.23	2.03	1.38	1.66 %	2.72 %	1.86 %
7 Hz	1.24	2.05	1.39	1.67 %	2.75 %	1.88 %
Original	1.29	2.15	1.42	1.72 %	2.87 %	1.90 %

As the approach most similar to a real world scenario is the '*Simulation*', Table 5.13 shows the MAE, RMSE and MAPE results for this approach, for windows with qualities '1' and '2', and for windows with quality '2'.

Table 5.13: Performance results for the '*Simulation*' approach for HR estimation.

High cutoff frequency	Both qualities			Quality '2'		
	MAE _{sim} (bpm)	RSME _{sim} (bpm)	MAPE _{sim}	MAE _{sim} (bpm)	RSME _{sim} (bpm)	MAPE _{sim}
3 Hz	4.88	11.45	5.84 %	1.40	2.71	1.89 %
4 Hz	4.67	11.14	5.59 %	1.34	2.38	1.81 %
5 Hz	4.53	10.87	5.43 %	1.37	2.40	1.84 %
6 Hz	4.57	10.90	5.51 %	1.38	2.44	1.86 %
7 Hz	4.55	10.79	5.48 %	1.39	2.45	1.88 %
Original	4.85	11.21	5.83 %	1.42	2.37	1.90 %

Figure 5.12 shows the Bland-Altman plot for the data filtered using the 0.5 Hz-5 Hz band-pass filter and the '*Simulation*' approach. The selection of the filter was based on the results presented on Table 5.13, on which this filter had the best performance when

using data with both qualities, the most difficult situation. The results using data with both qualities are presented on top with the mean difference between the two methods ($HR_{\text{estim}} - HR_{\text{ref}}$) being -1.51 beats/min (bpm) and the limits containing 95% of the data points being -1.51 ± 21.81 bpm. The results using only the best quality are presented at the bottom and the mean difference is 0.04 bpm, with the limits being 0.04 ± 4.65 bpm.

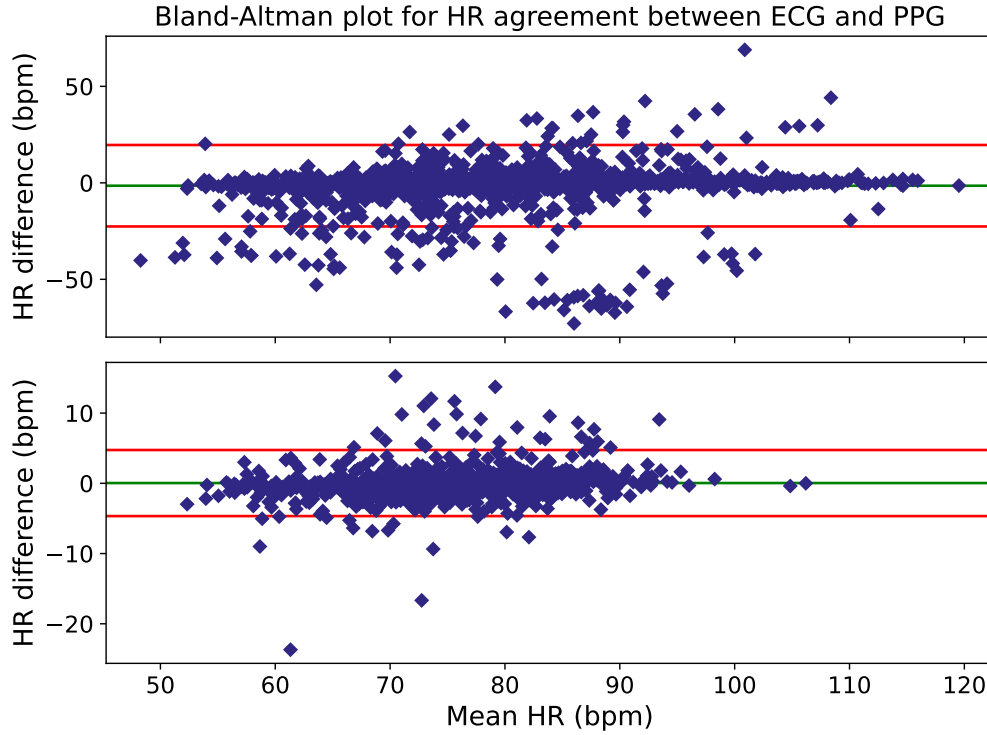


Figure 5.12: Bland-Altman plot for HR estimation, using the '*Simulation*' approach and the 0.5 Hz-5 Hz band-pass filter. The green line represents the mean difference between the two measurements and the red lines represent the limit for the 95% of the points ($\mu \pm 1.96\sigma$). On top are represented the estimates for windows with quality '1' and '2' (-1.50 ± 21.11). On the bottom are represented the estimates for windows with, at least, one channel with quality '2' (0.04 ± 4.70).

5.3.2 Respiratory Rate Extraction

This section presents the results from the RR extraction. This estimates were only calculated for windows with quality '2'. This decision was made given the fact that the algorithm to estimate this parameter is more sensitive to noise and artifacts, therefore, a higher quality signal is needed. Table 5.14 shows the MAE and MAPE for the three approaches used for RR estimation.

As the approach most similar to a real world situation is the '*Simulation*', the MAE, RMSE and MAPE results for every filter tested is shown in Table 5.15. There were two kinds of filters tested. One high-pass filter with a cutoff frequency of 0.05 Hz, two band-pass filters with cutoff frequencies of 0.08 Hz and 3 Hz and 0.05 Hz, and 5 Hz and the

Table 5.14: Performance results for the three approaches for the RR estimation.

Cutoff frequencies	MAE _{best} (brpm)	MAE _{avg} (brpm)	MAE _{sim} (brpm)	MAPE _{best}	MAPE _{avg}	MAPE _{sim}
0.05 Hz	4.14	4.26	4.26	28.13 %	29.11 %	29.11 %
0.08-3 Hz	3.43	3.40	3.40	23.19 %	23.31 %	23.31 %
0.05-5 Hz	4.38	4.45	4.45	29.23 %	29.75 %	29.75 %
Original	2.21	2.33	2.33	16.57 %	18.17 %	18.17 %

original filter, a digital filter composed by a band-pass with cutoff frequencies of 0.5 Hz and 10 Hz and a band-stop with cutoff frequencies of 40 Hz and 60 Hz.

Table 5.15: Performance results for the '*Simulation*' approach for the RR estimation.

Cutoff frequencies	MAE _{sim} (brpm)	RSME _{sim} (brpm)	MAPE _{sim}
0.05 Hz	4.26	5.21	29.11 %
0.08-3 Hz	3.40	4.20	23.31 %
0.05-5 Hz	4.45	5.46	29.75 %
Original	2.33	3.02	18.17 %

Figure 5.13 shows the Bland-Altman plot for the RR estimative with the original filter and the '*Simulation*' approach. This filter was chosen since it achieved the best results in Table 5.15. The average difference between the results calculated from the PPG signal and the RIP signal is -0.83 brpm and the limits containing 95% of the data points are -0.83 ± 5.69 brpm.

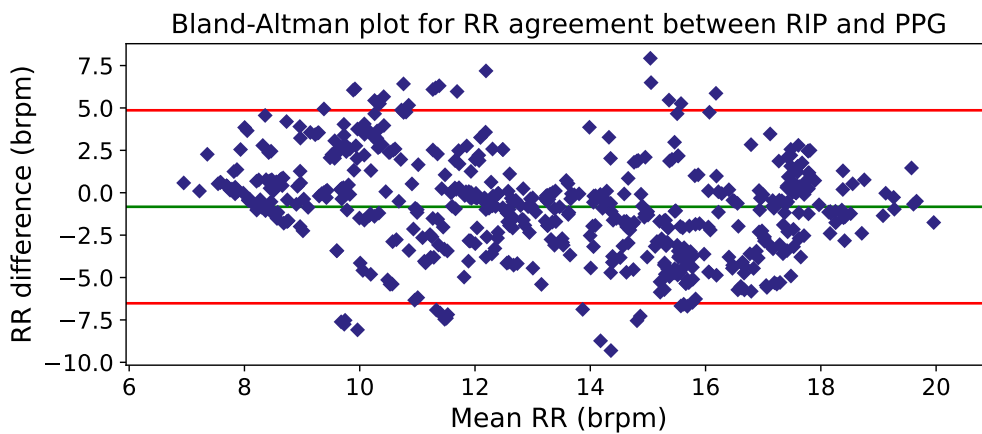


Figure 5.13: Bland-Altman plot for RR estimation, using the '*Simulation*' approach and the original filter. The green line represents the mean difference between the two measurements (-0.83 brpm) and the red lines represent the limit for the 95% of the points (-0.83 ± 5.69 brpm).

5.3.3 SpO₂ Extraction

This section presents the results from the SpO₂ extraction. The data used for this estimate was acquired with the acquisition protocol mentioned in Section 4.1.2, with tasks less prone to MA. Hence, the dataset is comprised of high quality data and it was only evaluated manually to exclude potential momentary MAs.

Since a linear regression model was used, each model was trained and tested with the same dataset, i.e., there was no division in train and test sets. However, a balancing step is present, therefore there is some variation on which data is used for each model. As a result, each model combination (tuning parameter λ - window size - R ratio expression) was trained 30 times for a better performance estimate.

Figure 5.14 shows the evolution of the RMSE with the increase of the tuning parameter λ . The different colors represent the different window sizes, in seconds, and each window has two plots representing the two expressions for the R ratio.

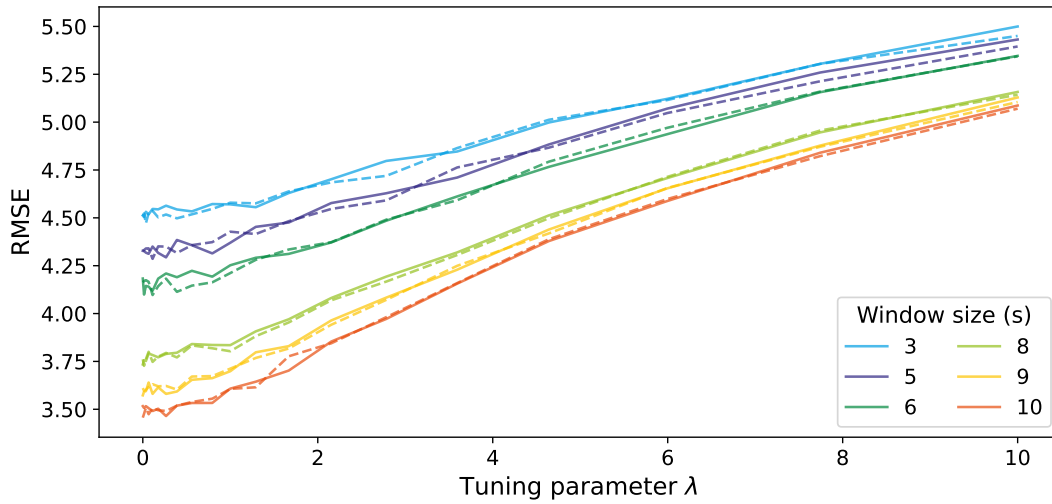


Figure 5.14: RMSE evolution with the increase of the tuning parameter λ . The different colors represent the different window sizes, in seconds, and the line style represents the two expressions for the R ratio, with the solid and dashed lines representing Equations 2.4 and 4.11, respectively.

Table 5.16 shows the results for the best tuning parameters for both R ratio expressions. These results were obtained using a 10 s window since, after an inspection of Figure 5.14 and the results obtained, a bigger window yields a more accurate result.

As can be seen in Table 5.16, the model with the lowest error used Equation 4.11 and $\lambda = 0.0$, hence, this model was used to present more detailed results. The average slope and y-intercept, along with their respective standard deviations, are presented in Equation 5.1. This transfer function is also represented in Figure 5.15, together with a balanced dataset. This dataset has MAE = 2.65 % SpO₂, RMSE = 3.41 % SpO₂, MAPE = 2.97 % and $R^2 = 0.73$.

Table 5.16: Results for the best SpO₂ extraction models. The top three tuning parameters are presented, for each expression used for the R ratio.

R ratio calculation	Tuning parameter λ	MAE (% SpO ₂)	RMSE (% SpO ₂)	MAPE (%)	R ²
Equation 2.4	0.27	2.63±0.13	3.47±0.14	2.95±0.14	0.72±0.02
	0.11	2.67±0.14	3.49±0.14	2.99±0.15	0.72±0.02
	0.18	2.68±0.16	3.50±0.16	3.00±0.17	0.72±0.03
Equation 4.11	0.00	2.64±0.12	3.46±0.12	2.95±0.13	0.73±0.02
	0.07	2.65±0.11	3.47±0.12	2.97±0.12	0.72±0.02
	0.01	2.69±0.11	3.48±0.12	3.01±0.12	0.72±0.02

$$SpO_2 = (-28.94 \pm 0.58) \times R + (112.43 \pm 0.50) \quad (5.1)$$

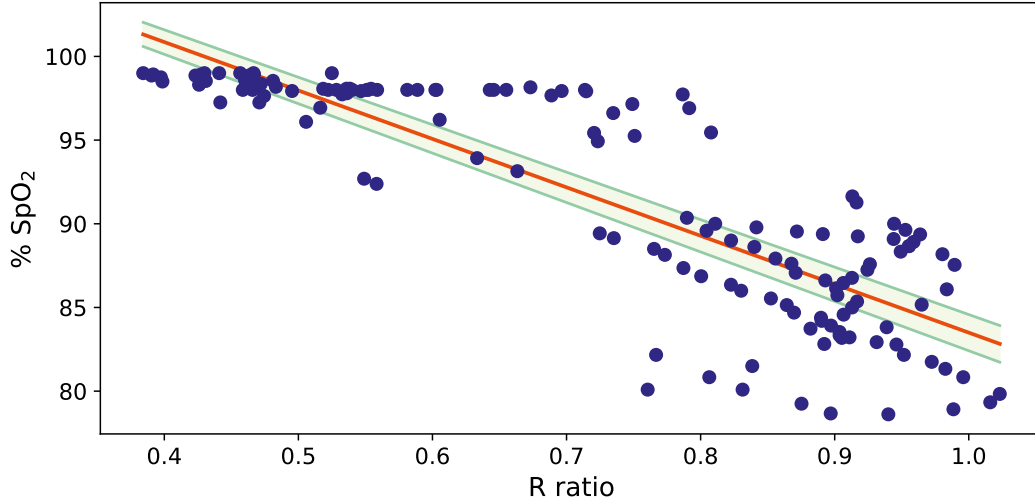


Figure 5.15: SpO₂ calibration curve for the proposed expression. The average linear function is represented in red with the error space represented in green.

Figure 5.16 shows the Bland-Altman for the SpO₂ estimate using the expression 5.1. The average difference between the results calculated with the expression 5.1 and the reference pulse oximeter is -0.03 % SpO₂ and the limits containing 95% of the data points are -0.03±6.71 % SpO₂.

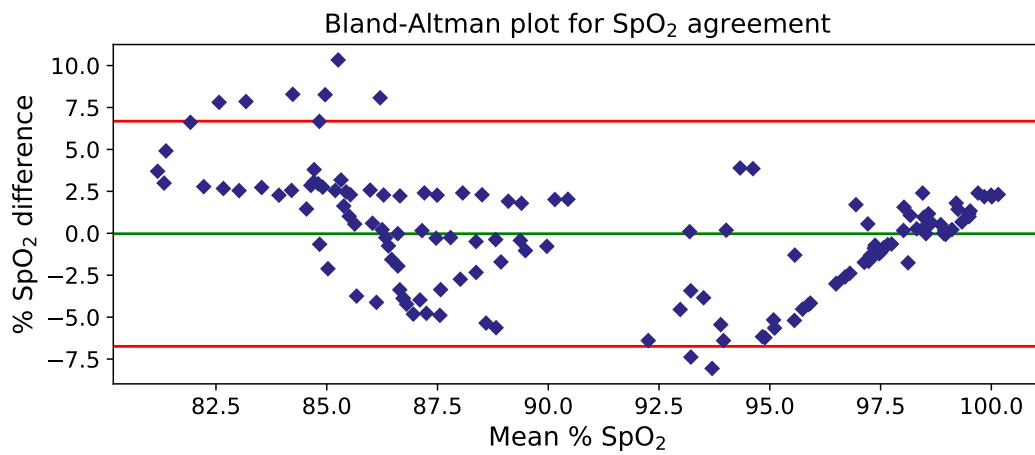


Figure 5.16: Bland-Altman plot for SpO₂ estimation. The green line represents the mean difference between the two measurements (-0.03% SpO₂) and the red lines represent the limit for the 95% of the points ($-0.03 \pm 6.71\%$ SpO₂).

Discussion

In this Chapter the results presented in Chapter 5 will be analyzed with further detail. Firstly, the results relating to the quality assessment will be discussed, including the different regression and classification models. Additionally, the extraction of the HR, RR and SpO₂ will also be evaluated.

6.1 Quality Assessment

As can be seen in Section 5.2, most of the data was manually classified as class '0', i.e., as corrupted signal. This result was expected since it is known that the PPG signal is highly susceptible to artifacts, as mentioned in Section 2.1.3. Table 5.2 also shows that the quality varies between the different subjects. This may be a result of a number of factors: (1) physiological elements, such as poor circulation; (2) more subject movement; (3) inadequate sensor positioning.

It is also possible to note that the red signal has more data with lower quality, with the class '0' representing 74.00% of the data in the red channel and representing 61.65% in the infrared channel.

6.1.1 Manual Classification Agreement

Section 5.2 also presents the results for the agreement between the two annotators. A Cohen's kappa of 0.813 and 0.971 were achieved in a 2 min portion of data for the red and infrared channels, respectively. These results show a very good, but not perfect, inter-annotator agreement. Furthermore, there could also exist some variance within each annotator, i.e., the same annotator can evaluate a segment in two different ways, as this process is highly subjective, even with guidelines [72].

These factors could result in discrepancies in the evaluations given to the ML models. Having all data manually classified by all annotators and having more annotators would help to mitigate these effects. This would result in a more precise evaluation that would improve the dataset quality. Nevertheless, these results show that there is a high agreement between annotators and that the data manually evaluated can be used for further

development.

6.1.2 Regression Models

Table 5.3 presents the performance results for the different regression models tested. It is possible to note that the majority of the models do not perform well in differentiating between the different qualities, which can be seen by the negative R^2 values. The best models are the Random Forest and Gradient Boost. They are the only ones with a positive R^2 and have very similar results. The RF regressor was chosen for test for further optimization.

Figure 5.2 shows the regression results, in the form of RMSE, for the red and infrared channels in function of the number of trees generated and the maximum depth. It is possible to note that the behavior for the different combinations are similar, even though the values may differ. As expected, the error decreases with the increase of the number of trees generated until an approximate stable value. Figure 5.2 also shows that the error increases with the increase in the maximum depth allowed, after an initial drop. This behavior was expected since a deep tree can have difficulty generalizing, as mentioned in Section 2.3.1.

Table 5.4 shows the results for the best combinations for the red and IR models and Figure 5.3 shows the true and predicted values for the red channel regressor. It was expected that, for a good regression model, a diagonal line was displayed between the true and predicted values. Instead, the ML model predicts almost all data points as qualities between 0.25 and 1 and fails to accurately predict the real signal quality.

Approaches using a regression model for PPG quality assessment were not found in the literature, therefore, these results can not be compared to anything. Nevertheless, the outcomes were not satisfactory and, thus, no further examination or improvements were done. These poor results could be the effect of several factors: (1) inadequate signal features; (2) limited available data; (3) few intermediate values between the three main qualities. As the data was manually classified with three qualities, the majority of the windows will have one of these three qualities, since there can only be decimal values in the boundaries of the manual classification.

6.1.3 Classification Models

Table 5.5 presents the results for the different types of classifiers initially tested. It is possible to note that only the decision tree, the Random Forest and the Gaussian Naive Bayes classifiers perform with an accuracy better than random choice, with the RF classifier having the best results. This model was chosen for further optimization, by testing different parameters.

Two distinct cases were studied: (1) using features from both channels to train both classifiers and (2) training each channel classifier with features only from the respective channel. When analyzing Figures 5.4 and 5.6, it is possible to note that the performance

behavior is very similar for the different models when increasing the number of trees generated or when increasing the maximum depth. As expected, the accuracy increases with the number of trees, until a plateau. With the increase on maximum depth, the accuracy has an initial peak in performance, then, stabilizing with a lower accuracy. The value of this peak varies for the different cases, with the classifiers for the red channel benefiting more with lower depths, when comparing with the IR classifiers. This behavior can be expected since a deep tree can have difficulty in generalizing, as mentioned in Section 2.3.1. It is also possible to note that the red channel classifiers have more pronounced peaks in performance, when compared with the IR classifiers, especially when they are only trained with features from individual channels.

Tables 5.6 and 5.7 present the results for the best classifier when both channels features are used and when individual features are used, respectively. Comparing these results, models trained with all features perform better classifying the red channel quality, as opposed to classifiers with individual features, which perform better on classifying the infrared channel quality. This can indicate that the models for the red channel quality rely on the infrared features for better results and the IR classifiers are disturbed by the features of the red signal. This suggestion can also be observed in Figure 5.10, where the top 4 correlated features with the red quality are from the IR channel. This may also suggest that the red channel has low quality when the IR has low quality, but the opposite does not occur. In tables 5.6 and 5.7 it is also possible to observe that the '*Average*' quality has better results than the '*Mode*' quality interpretation, even though these differences are contained in one standard deviation error.

Figures 5.5 and 5.7 show the confusion matrices for the best models, using features from all channels and from individual channels, respectively. These matrices give a different perspective to the results gathered. Figure 5.5 shows that the models using the '*Average*' quality perform similarly to the ones using the '*Mode*' quality. It is also possible to note that all models can accurately differentiate between classes '0' and '2', but can struggle to differentiate between classes '0' and '1', and classes '1' and '2'. These results were expected since there is a higher similarity between adjacent classes (classes '0' and '1', and '1' and '2') which results in an increased difficulty in differentiating between them. In contrast, classes '0' and '2' have a lower similarity, resulting in a better distinction between them. These findings lead to the development of double classifier.

Figure 5.7 also shows that the IR classifier performs very similarly when using only features from that channel. In contrast, the red classifier has a great difficulty discriminating between the different classes when using only features from the respective channel.

As mentioned in Section 3.1.3, most PPG quality assessment research focus on binary classification, hence, comparative results to the multi-class approach used in this work are limited. Moscato et al. [115] and Elgendi [76] divided the signal into three qualities, similar to the developed in this work, but the classifiers were developed with a concept similar to the double classifier, using two binary classifiers. Pradhan, Rajan, and Adler [72] divided the signal into five qualities and used wrist-based PPG. The authors

used a subject-wise cross-validation and their best model, a Random Forest classifier trained with PPG and accelerometer features, achieved an accuracy of 74.5%. This result is slightly better than the one achieved in this work. However, the results achieved in this work are still promising. The models developed achieved an accuracy double than random selection with a simple and fast algorithm that only uses features from a single sensor, thus giving it independence, i.e., contrary to the reported work that use accelerometer features, our algorithms do not rely on any other sensors.

Nevertheless some improvements could be made to improve the performance attained. As with the regression models, these results could be improved with the study of better signal features and a larger dataset.

6.1.3.1 Double Classifier

An alternative studied was the development of a double classifier, i.e., one classifier to separate the data between class '0' and ' $1 \cup 2$ ' classes, and a second classifier to evaluate between the classes '1' and '2'. This separation was based on the fact that physiological information can be extracted from the classes '1' and '2', unlike the class '0'. Additionally, the multi-class classifiers showed some difficulty in differentiating between the classes '1' and '2'. Hence, two binary classifiers were developed.

Figure 5.8(a) shows the confusion matrices for the best models for classifying between classes '0' and ' $1 \cup 2$ '. As expected by the results achieved in the multi-class classifiers, these models can accurately differentiate between classes '0' and ' $1 \cup 2$ '. These outcomes are also confirmed by the results in Table 5.8, with the red and IR classifiers having, approximately, 80% accuracy, for both quality interpretations.

Figure 5.8(b) presents the confusion matrices for the best models for classifying between classes '1' and '2'. It is possible to see that the infrared channel classifiers perform better than the red channel classifiers, which is also shown in Table 5.9. As stated in Section 6.1.3, the red channel classifiers could be relying on the infrared features for better results. This could result in a situation where the red channel has a different quality than the IR channel, consequently, the red channel classifier would have difficulty in correctly classifying the quality of the corresponding channel.

It is also possible to note that, when classifying between the classes '1' and '2', the IR channel classifiers perform better using the '*Mode*' quality and, by contrast, the red channel classifiers have a better performance when using the '*Average*' quality interpretation.

One important aspect to consider when using chained classifiers is the error propagation [115], i.e., a window with class '0' could be wrongly categorized in the first classifier propagating this error to the second classifier, therefore the real-world performance could be worse than the one stated, since the reported models were only trained and tested with the real classes.

Moscato et al. [115] tested two strategies for differentiating between three qualities: (1) a classifier to detect classes '0' and ' $1 \cup 2$ ' and a classifier to detect classes '2' and ' $0 \cup 1$ ';

(2) a chained classifier, similar to the approach developed in this work. The authors achieved an accuracy of 95 % for the classifier between classes '0' and ' $1 \cup 2$ ', an accuracy of 96 % for the classifier between classes '2' and ' $1 \cup 2$ ', and an accuracy of 93 % for the classifier between classes '1' and '2'. Moscato et al. [115] also reported the results achieved by Elgendi [76] with the accuracies being 72 %, 78 % and 59 %, respectively. The classifiers developed in this work did not perform as well as the ones developed by Moscato et al. [115] but, in contrast to this thesis, the authors used features from the accelerometer signal, did not balance their dataset and did not use a subject-wise cross-validation, which is reported to provide a more correct performance evaluation for a real world scenario [51]. By contrast, the classifiers developed by Elgendi [76], which use a single SQL, had worst results than the one developed in this thesis.

Since there is information that can be extracted from a signal with a quality belonging to either class '1' or '2', a classifier with 80 % accuracy, as the ones developed in this work, could be helpful in a wearable device. Additionally, this type of chained classifier could be a possible solution for further improvement, e.g., developing specific features for differentiating between the classes '1' and '2'.

6.1.3.2 Vote Classifiers

Another alternative studied was the use of three different classifiers that vote on a final classification. For each channel and quality interpretation the best RF model was used in addition to a Gaussian Naive Bayes and a single decision tree.

Table 5.10 shows the performance results for the best combinations of the vote classifiers. There is no significant improvement in relation to the results shown in Table 5.6, with the results falling in the error interval. Figure 5.9 also shows the confusion matrices for these models. Again, the results for the vote classifiers did not improve the results achieved by a single Random Forest classifier. Hence, it is preferred to use a simpler model, as stated in Occam's razor [116]. When tested initially, these additional classifiers had a lower performance compared to RF model, as shown in Table 5.5. Therefore, the Random Forest model already correctly classifies what these new models classify.

Consequently, a possible improvement could be the use of a different set of classifiers, in particular, a set of classifiers that complement each other, i.e., a set of models that classify different samples in different ways. Additionally, a bigger collection of models could also improve the results. These suggestions could be computationally more expensive, therefore, a balance between the models performance and the device capability may be needed.

6.1.4 Feature Importance

Section 5.2.3 presents the relationship between the quality of each signal and their features, in form of the Pearson correlation. Figure 5.10 presents the correlation between all features and the quality from the two channels, for both quality interpretations. It

is possible to note that the order by which the features are presented is the same for the 'Average' and 'Mode' qualities, despite existing some variation on the correlation value.

It is also possible to note that the top four correlated features with the red channel quality are from the IR channel. This effect, once more, indicates that the red channel classifiers use IR information for better results, which was already discussed in Section 6.1.3. Figure 5.10 also shows that the features least correlated with the IR channel quality belong to the red channel. This may indicate that the IR classifiers that use all features can be disturbed by these features.

Figures 5.11(a) and 5.11(b) present the accuracy for the models trained with different amounts of features, for the red and infrared channels respectively. It is possible to note that red channel classifiers performance increases with the increase of features. By contrast, the IR classifiers performance decreases when more features from the red channel are used. It is also possible to note that the IR models trained with more than 10 features have a bigger difference between them, represented by a bigger error bar. Once again, it indicates that the IR classifiers are disturbed by the red channel features.

A further investigation could be made in order to find which feature combination is optimal for each channel classifier. There are several methods that could be used for features selection, e.g., Boruta algorithm, which uses several tests with random forest models to identify important and unimportant features [43]. A smaller selection of features could potentially decrease the computational power needed and possibly lead to an improvement of the models' accuracy.

6.2 Physiological Metrics

6.2.1 Heart Rate Extraction

Tables 5.11 and 5.12 show the results for the different approaches for the HR calculation. The 'Best' approach had the best results. This was an expected result since, with this approach, the algorithm always chooses the channel with the highest quality to use and, since there is no rounding of the window quality, the quality is a more precise number than with the 'Simulation' approach. The 'Average' approach has the worst results as it uses both channels, where one of them could have a worse quality, deteriorating the final estimate. The 'Simulation' had an intermediate performance, as expected, since this approach could be seen as an hybrid between the two other approaches.

It is also possible to notice that the dataset with, at least, one channel classified as quality '2' had the best results, when comparing with the data with one channel with quality '1' or better. This was anticipated since data with more quality is expected to give more accurate results.

As the 'Simulation' approach is the most approximate to a real world scenario, the MAE, RMSE and MAPE for this concept can be seen in Table 5.13. It is possible to note that a stricter filter can improve the results. This can be due to the fact that a filter with

lower cutoff frequencies can remove higher frequency noise, hence remove false peaks that would otherwise be detected.

The results using both qualities can be compared with the results from commercially available products such as the Apple Watch 3 and Fitbit Charge 2 [6] and can be considered accurate results as they have less than 10% of error [6, 113]. Nelson and Allen [6] also reported that some health science research only consider an acceptable error of 5%. The results using both qualities do not comply with these more rigorous standards, with the minimum MAPE achieved being 5.4%. The results using data with, at least, one channel with quality '2' adhere to the more precise errors, with the minimum error achieved being 1.8%.

As can be seen in Figure 5.12, there is no visible systematic under or over estimation of the HR. Thus, the existent errors can arise from two main sources: (1) the algorithm detects false peaks in the PPG signal, increasing the HR_{estm} , or has problems detecting some peaks, decreasing the HR_{estm} . This can be a result of a misclassified quality, where the actual quality is lower than the one given to the algorithm; (2) the algorithm does not detect correctly the peaks in the ECG signal. Since this signal was not evaluated for its quality, even though it is more resistant to MA, it can have segments with low quality, hindering the HR_{ref} . This effect was visually inspected on some estimates.

6.2.2 Respiratory Rate Extraction

Similar to the HR extraction, Table 5.14 shows that the '*Best*' approach yields the best results. The '*Average*' and '*Simulation*' approach had the same results because they use the same data. This happens since the RR is only extracted when, at least, one channel has quality '2'. Therefore, when only one channel meet the criteria, the RR_{estm} is the estimate for that channel. When both channels meet the criteria, the RR_{estm} is the average from both channels. This happens for both approaches. The '*Best*' approach has slightly different results because it does not use the rounded window quality, hence it never averages the results and only uses the highest quality window.

As the '*Simulation*' approach is the most approximate to a real world scenario, the MAE, RMSE and MAPE for this concept can be seen in Table 5.15. It is possible to note that the original filter (band-pass with cutoff frequencies of 0.5 Hz and 10 Hz followed by a band-stop with cutoff frequencies of 40 Hz and 60 Hz) gave the best results. These results, were similar to the article on which the developed algorithm was based, with a RMSE of ≈ 3 brpm which they considered promising [69]. The MAE is also comparable with the results achieved by Pimentel et al. [70], in which they were able to estimate the RR with a MAE of 1.5 and 4.0 brpm, depending on the dataset used. The proposed algorithm also has some tendency to underestimate the RR, shown as a negative mean difference of -0.8 brpm in Figure 5.13.

As mentioned in Section 4.5.2, multiple studies use capnometry as the reference method for RR estimation. This was not possible in this study and a RIP sensor was

used. Thus, even though the results presented are promising, a validation with another reference method and more data is needed before clinical use.

6.2.3 SpO₂ Extraction

Section 5.3.3 presents the results from the SpO₂ extraction. As several liner models were tested, Figure 5.14 shows the RMSE with the increase in the tuning parameter λ , for the different window sizes used. It is possible to see that a bigger λ results in a higher error, making the preferred Ridge model closer to an Ordinary Least Squares regression. It is also possible to note that with an increase in the window size, there is a decrease in the RMSE. A bigger window size results in an average of more R ratios, since this is calculated for every pulse, hence, an abnormal pulse has less influence on the final SpO₂ value.

Table 5.16 shows the best models for extracting SpO₂ from the R ratio, using a 10 s window. As expected, the optimal λ has a small value. It is also possible to note that the models using both R ratio expressions have very similar results, with all the variation inside the error margin. The model with lower RMSE used Equation 4.11 and a tuning parameter of $\lambda = 0$ and achieved a MAE = 2.64 % SpO₂, RMSE = 3.46 % SpO₂, MAPE = 2.95 % and $R^2 = 0.73$. These results comply with the ISO standards which states that a pulse oximeter should have a RMSE < 4.0 % [54].

In Equation 5.1 it is possible to see the final transfer function between the R ratio and the SpO₂ value, and Figure 5.15 shows this expression, along with a balanced dataset. This equation was produced as the average of the different models with the same parameters (tuning parameter λ - window size - R ratio expression), and the error limits represent the standard deviation. While not perfect, it is possible to see a linear relationship between the R ratio and the % SpO₂ value. Figure 5.16 shows the Bland-Altman plot for these balanced dataset. It is possible to see that there is no systematic under or over estimation by the proposed method, given that the average difference between the expression achieved and the reference method is -0.03 % SpO₂.

While these results are promising, there are some issues with the acquired data: (1) the data was only extracted from one subject; (2) the reference SpO₂ was mainly contained in the 80-100 % SpO₂ range. As established in the ISO 80601-2-61 [54], the protocol should include a sufficient number of subjects, healthy adults both males and females, to reach statistical significance. Therefore, more subjects with a wider SpO₂ range is needed for a more accurate measurement.

Conclusion

With the advancement of technology, wearable devices have been promoted and improved in the last years. In addition, their application in the digital measurement of health has gained attention by researchers, as they allow for continuous data acquisition in real-world scenarios, however, it could be at the cost of the signal quality. A signal with low quality, often corrupted by artifacts, can be impossible to interpret, however, a lightly corrupted PPG signal can still provide physiological information, such as the heart rate. Therefore, one of this thesis objectives was the development of a solution that can detect the signal quality in real-time. The solution developed divided the data into three separate qualities (optimal, suboptimal, corrupted) for enabling a distinct processing whether the signal has sufficient quality for extracting all, some or none biological metrics.

During the development of this thesis, a data acquisition protocol and an annotated database were also created, which can be used in the future. The protocol focused on acquiring data during daily activities and the sensor was positioned to mimic a smart-watch placement. This resulted in a dataset with different qualities that was subsequently manually evaluated. This evaluation was the base for the ML models, whether regression or classification, given that a supervised learning approach was used for an automatic quality assessment stage. A second data acquisition protocol was also developed with its main focus being the extraction of the SpO₂ value. This protocol had tasks less prone to motion artifacts and an exercise for allowing the subject to be in apnea for longer periods of time than in normal circumstances.

A regression approach was initially tested for a continuous, hence, more precise value of the window quality, however, the results were not satisfactory. Therefore, several classification models were developed. The multi-class classifiers to identify the three signal quality levels achieved an accuracy double than random chance, similar to other systems found in the literature. A chained classifier was also tested, where the system initially detects whether the signal has sufficient quality for extracting biological metrics, and then differentiates between the best qualities, i.e., the quality that allows to extract all metrics and the quality that allows to extract some identified metrics but not all, namely, the SpO₂ value. These models also had an adequate performance, specially, the first

step of the system. A feature selection stage was also studied and showed that a smaller feature set can improve the classifiers performance, however, a more in-depth study is needed for a better understanding on the ideal set of features.

The heart rate and respiratory rate were also extracted from the PPG signal. Since there is a prior signal quality evaluation, these metrics are only extracted when the quality exceeds a threshold, thus, avoiding abnormal values. Both algorithms developed resulted in performances similar to those found in the literature and in other devices currently on the market. A SpO₂ extraction algorithm was also developed. While the results achieved are promising, more data is needed to reach statistical significance.

7.1 Future Work

While this work presents promising results, there are still improvements that could be made before applying the developed algorithms in a real-world device.

- **Expanding the database**

One of the biggest improvements possible would be a bigger population. In this work, data from five subjects were acquired and only data from three subjects could be used for ML development. A bigger sample size would provide a better statistical significance, while evaluating more correctly the models ability to generalize to new subjects. The current database could also be improved with a bigger sample size of signals with classes '1' and '2'. Additionally, the models could be trained and tested in subjects with arrhythmias or other cardiac pathologies, as the models developed in this study would not been able to discriminate between arrhythmias and noise.

- **Adding annotators**

Another possible improvement would be a bigger pool of annotators. A larger group of annotators would allow the same segment being evaluated by more than one person, improving the quality of the annotations, since this process can be subjective [72].

- **Feature engineering**

A more in-depth study of the features extracted from the signal could improve significantly the results. There are several features used by different studies, such as, correlation between consecutive pulses [115], zero-crossing rate from the signal first derivative [73, 115] and similarity between the second derivative of adjacent pulses [117]. Another alternative could be the implementation of features from other sensors, e.g., the accelerometer. This leads to a loss in independence but could improve the results.

- **Filtering**

Since the metrics calculated in this work do not change abruptly, smoothing filters could also be applied to avoid abrupt changes in the value. Kalman filters have been applied in SpO₂ extraction [111] and RR extraction [69, 70]. This type of filter can also be used for removing MA in the PPG signal [79].

Bibliography

- [1] J. M. Lourenço. *The NOVAtesis L^AT_EX Template User's Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/master/template.pdf>.
- [2] W. R. Thompson. "Worldwide Survey of Fitness Trends for 2022". In: *ACSM's Health and Fitness Journal* 26.1 (Jan. 2022), pp. 11–20. ISSN: 1536593X. DOI: 10.1249/FIT.0000000000000732.
- [3] J. Allen and P. A. Kyriacou. *Photoplethysmography: Technology, Signal Analysis and Applications*. Elsevier, 2021.
- [4] A. R. Watson, R. Wah, and R. Thamman. "The value of remote monitoring for the COVID-19 pandemic". In: *Telemedicine and e-Health* 26.9 (Sept. 2020), pp. 1110–1112. ISSN: 15563669. DOI: 10.1089/tmj.2020.0134.
- [5] K. T. Sweeney, T. E. Ward, and S. F. McLoone. "Artifact removal in physiological signals-practices and possibilities". In: *IEEE Transactions on Information Technology in Biomedicine* 16.3 (May 2012), pp. 488–500. ISSN: 10897771. DOI: 10.1109/TITB.2012.2188536.
- [6] B. W. Nelson and N. B. Allen. "Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study". In: *JMIR mHealth and uHealth* 7.3 (Mar. 2019), e10828. ISSN: 22915222. DOI: 10.2196/10828.
- [7] J. P. Buschmann and J. Huang. "New ear sensor for mobile, continuous and long term pulse oximetry". In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*. IEEE, Aug. 2010, pp. 5780–5783. ISBN: 9781424441235. DOI: 10.1109/IEMBS.2010.5627835.
- [8] L. Smital et al. "Real-Time Quality Assessment of Long-Term ECG Signals Recorded by Wearables in Free-Living Conditions". In: *IEEE Transactions on Biomedical Engineering* 67.10 (2020), pp. 2721–2734. DOI: 10.1109/TBME.2020.2969719.

- [9] Y. Shi et al. "Robust Assessment of ECG Signal Quality for Wearable Devices". In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. 2019, pp. 1–3. DOI: 10.1109/ICHI.2019.8904810.
- [10] A. Henriksen et al. "Using fitness trackers and smartwatches to measure physical activity in research: Analysis of consumer wrist-worn wearables". In: *Journal of Medical Internet Research* 20.3 (Mar. 2018), e110. ISSN: 14388871. DOI: 10.2196/jmir.9157.
- [11] T. Tamura. "Current progress of photoplethysmography and SPO2 for health monitoring". In: *Biomedical Engineering Letters* 9.1 (Feb. 2019), pp. 21–36. ISSN: 2093985X. DOI: 10.1007/s13534-019-00097-w.
- [12] K. D. Torp, P. Modi, and L. V. Simon. *Pulse oximetry*. Accessed: 09-02-2022. Feb. 2021. URL: <https://www.ncbi.nlm.nih.gov/books/NBK470348/>.
- [13] A. W. Gelb et al. "World Health Organization-World Federation of Societies of Anaesthesiologists (WHO-WFSA) international standards for a safe practice of anesthesia". In: *Anesthesia and Analgesia* 126.6 (June 2018), pp. 2047–2055. ISSN: 15267598. DOI: 10.1213/ANE.0000000000002927.
- [14] J. Tang, A. Mandrusiak, and T. Russell. "The feasibility and validity of a remote pulse oximetry system for pulmonary rehabilitation: A pilot study". In: *International Journal of Telemedicine and Applications* 2012 (2012), pp. 1–7. ISSN: 16876415. DOI: 10.1155/2012/798791.
- [15] F. Braun et al. "Pulse Oximetry at the Wrist during Sleep: Performance, Challenges and Perspectives". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2020-July* (July 2020), pp. 5115–5118. ISSN: 1557170X. DOI: 10.1109/EMBC44109.2020.9176081.
- [16] P. J. Chacon et al. "A Wearable Pulse Oximeter With Wireless Communication and Motion Artifact Tailoring for Continuous Use". In: *IEEE Transactions on Biomedical Engineering* 66.6 (June 2019), pp. 1505–1513. ISSN: 15582531. DOI: 10.1109/TBME.2018.2874885.
- [17] G. Priem et al. "Clinical grade SpO2 prediction through semi-supervised learning". In: *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*. IEEE, Oct. 2020, pp. 914–921. ISBN: 9781728195742. DOI: 10.1109/BIBE50027.2020.00155.
- [18] P. Prasun, S. Mukhopadhyay, and R. Gupta. "Real-time multi-class signal quality assessment of photoplethysmography using machine learning technique". In: *Measurement Science and Technology* 33.1 (Jan. 2022), p. 015701. ISSN: 13616501. DOI: 10.1088/1361-6501/ac2d5b.
- [19] *PubMed*. Accessed: 31-08-2022. URL: <https://pubmed.ncbi.nlm.nih.gov/>.

-
- [20] M. Nitzan, A. Romem, and R. Koppel. "Pulse oximetry: Fundamentals and technology update". In: *Medical Devices: Evidence and Research* 7.1 (July 2014), pp. 231–239. ISSN: 11791470. DOI: 10.2147/MDER.S47319.
 - [21] M. Sauthier et al. "Estimated Pao 2: A Continuous and Noninvasive Method to Estimate Pao 2 and Oxygenation Index". In: *Critical Care Explorations* 3.10 (Sept. 2021), e0546. ISSN: 2639-8028. DOI: 10.1097/cce.0000000000000546.
 - [22] T. Aoyagi. "Pulse oximetry: Its invention, theory, and future". In: *Journal of Anesthesia* 17.4 (Nov. 2003), pp. 259–266. ISSN: 09138668. DOI: 10.1007/s00540-003-0192-6. URL: <http://link.springer.com/10.1007/s00540-003-0192-6>.
 - [23] S. Alharbi et al. "Oxygen saturation measurements from green and orange illuminations of multi-wavelength optoelectronic patch sensors". In: *Sensors (Switzerland)* 19.1 (Dec. 2019), p. 118. ISSN: 14248220. DOI: 10.3390/s19010118.
 - [24] T. Tamura et al. "Wearable photoplethysmographic sensors—past and present". In: *Electronics* 3.2 (Apr. 2014), pp. 282–302. ISSN: 20799292. DOI: 10.3390/electronics3020282.
 - [25] P. A. Kyriacou. "Pulse oximetry in the oesophagus". In: *Physiological Measurement* 27.1 (Jan. 2006). ISSN: 13616579. DOI: 10.1088/0967-3334/27/1/R01.
 - [26] A. P. Mohamed Tanveejul et al. "A Study on the Subject and Location Specificity in Reflectance based SpO2 Estimation using R-value based Calibration Curve". In: *IEEE Medical Measurements and Applications, MeMeA 2020 - Conference Proceedings*. IEEE, June 2020, pp. 1–6. ISBN: 9781728153865. DOI: 10.1109/MeMeA49120.2020.9137227.
 - [27] *Optical absorption of hemoglobin*. Accessed: 28-11-2021. URL: <https://omlc.org/spectra/hemoglobin/>.
 - [28] O. Y. Hay et al. "Pulse oximetry with two infrared wavelengths without calibration in extracted arterial blood". In: *Sensors (Switzerland)* 18.10 (Oct. 2018), p. 3457. ISSN: 14248220. DOI: 10.3390/s18103457.
 - [29] A. Von Chong et al. "Towards a novel single-LED pulse oximeter based on a multispectral sensor for IoT applications". In: *Microelectronics Journal* 88 (June 2019), pp. 128–136. ISSN: 00262692. DOI: 10.1016/j.mejo.2018.03.005.
 - [30] K. Budidha, V. Rybynok, and P. A. Kyriacou. "Design and Development of a Modular, Multichannel Photoplethysmography System". In: *IEEE Transactions on Instrumentation and Measurement* 67.8 (Aug. 2018), pp. 1954–1965. ISSN: 00189456. DOI: 10.1109/TIM.2018.2810643.
 - [31] T. Esmaeiliazad. "Special Study of Earlobe Pulse Oximeter Using MAX30100 for Detecting SpO2 and Heart Beat". In: *Academic Journal of Health Sciencies: Medicina balear*, ISSN-e 2255-0560, Vol. 36, N°. 4, 2021, págs. 43-47 36.4 (2021), pp. 43–47. ISSN: 2255-0560. DOI: 10.3306/AJHS.2021.36.04.43.

- [32] M. T. Petterson, V. L. Begnoche, and J. M. Graybeal. “The effect of motion on pulse oximetry and its clinical significance”. In: *Anesthesia and Analgesia* 105.SUPPL. 6 (Dec. 2007), S78–S84. issn: 00032999. doi: 10.1213/01.ane.0000278134.47777.a5.
- [33] D. Seok et al. “Motion Artifact Removal Techniques for Wearable EEG and PPG Sensor Systems”. In: *Frontiers in Electronics* 2 (May 2021). issn: 2673-5857. doi: 10.3389/felec.2021.685513.
- [34] G. W. J. Clarke. “Signal Quality Analysis in Pulse Oximetry: Modelling and Detection of Motion Artifact”. MA thesis. Carleton University, 2015.
- [35] J. Yao and S. Warren. “A Short Study to Assess the Potential of Independent Component Analysis for Motion Artifact Separation in Wearable Pulse Oximeter Signals”. In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. 2005, pp. 3585–3588. doi: 10.1109/IEMBS.2005.1617256.
- [36] *Top 10 Health Technology Hazards for 2015*. ECRI Institute, Nov. 2014.
- [37] L. Wiklund et al. “Postanesthesia monitoring revisited: Frequency of true and false alarms from different monitoring devices”. In: *Journal of Clinical Anesthesia* 6.3 (1994), pp. 182–188. issn: 0952-8180. doi: 10.1016/0952-8180(94)90056-6.
- [38] J. Fletcher, M. Page, and H. E. Jeffery. “Sleep States and Neonatal Pulse Oximetry”. In: *Sleep* 21.3 (May 1998), pp. 305–310. issn: 0161-8105. doi: 10.1093/sleep/21.3.305.
- [39] P. H. Charlton et al. “Acquiring Wearable Photoplethysmography Data in Daily Life: The PPG Diary Pilot Study”. In: *Proceedings of 7th International Electronic Conference on Sensors and Applications*. Basel, Switzerland: MDPI, Nov. 2020, p. 80. doi: 10.3390/ecsa-7-08233.
- [40] N. Stubán and N. Masatsugu. “Non-invasive calibration method for pulse oximeters”. In: *Periodica Polytechnica Electrical Engineering* 52.1-2 (2008), p. 91. issn: 0031-532X. doi: 10.3311/pp.ee.2008-1-2.11.
- [41] G. Bonaccorso. *Machine learning algorithms: Reference guide for popular algorithms for Data Science and Machine Learning*. Packt Publishing, 2017.
- [42] E. Alpaydin. *Machine Learning*. MIT Press, 2021.
- [43] A. Burkov. *Machine Learning Engineering*. 1st ed. 2022.
- [44] G. Dong and H. Liu. *Feature Engineering for Machine Learning and Data Analytics*. CRC Press, 2018.
- [45] B. Boehmke and B. Greenwell. *Hands-on machine learning with R*. 1st ed. Chapman and Hall/CRC, 2019.
- [46] M. I. L. das Neves. “Opening the black-box of artificial intelligence predictions on clinical decision support systems”. MA thesis. NOVA School of Science and Technology, 2020.

-
- [47] E. Frank and I. H. Witten. "Selecting multiway splits in decision trees". In: 1996.
 - [48] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
 - [49] I. H. Witten and E. Frank. *Data mining.practical machine learning tools and techniques*. 2nd ed. Elsevier, 2005.
 - [50] D. Berrar. "Cross-validation". In: *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Vol. 1-3. Elsevier, 2018, pp. 542–545. ISBN: 9780128114322. DOI: 10.1016/B978-0-12-809633-8.20349-X. arXiv: 1703.03167.
 - [51] S. Saeb et al. "Voodoo Machine Learning for Clinical Predictions". In: *bioRxiv* (2016). DOI: 10.1101/059774.
 - [52] H. M and S. M.N. "A Review on Evaluation Metrics for Data Classification Evaluations". In: *International Journal of Data Mining & Knowledge Management Process* 5.2 (Mar. 2015), pp. 01–11. ISSN: 2231007X. DOI: 10.5121/ijdkp.2015.5201.
 - [53] M. Z. Naser and A. H. Alavi. "Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences". In: *Architecture, Structures and Construction* (Nov. 2021). ISSN: 2730-9886. DOI: 10.1007/s44150-021-00015-8.
 - [54] *Medical electrical equipment – Particular requirements for basic safety and essential performance of pulse oximeter equipment*. ISO 80601-2-61:2017. Dec. 2017.
 - [55] D. Chicco, M. J. Warrens, and G. Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation". In: *PeerJ Computer Science* 7 (July 2021), e623. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.623.
 - [56] P. Domingos. *The master algorithm: how the quest for the ultimate learning machine will remake our world*. Penguin Books, 2017.
 - [57] J. Brownlee. *Imbalanced Classification with Python*. 1st ed. Self-published, 2021.
 - [58] S. Hossain, T. H. Kwon, and K. D. Kim. "Comparison of Different Wavelengths for Estimating SpO2 Using Beer-Lambert Law and Photon Diffusion in PPG". In: *ICTC 2019 - 10th International Conference on ICT Convergence: ICT Convergence Leading the Autonomous Future*. IEEE, Oct. 2019, pp. 1377–1379. ISBN: 9781728108926. DOI: 10.1109/ICTC46691.2019.8939849.
 - [59] P. P. Banik et al. "Development of a wearable reflection-type pulse oximeter system to acquire clean ppg signals and measure pulse rate and spo2 with and without finger motion". In: *Electronics (Switzerland)* 9.11 (Nov. 2020), pp. 1–26. ISSN: 20799292. DOI: 10.3390/electronics9111905.
 - [60] V. Vizbara. "Comparison of Green, Blue and Infrared Light in Wrist and Forehead Photoplethysmography". In: *BioMed* 2013. 2013.

- [61] A. Kiruthiga et al. "Reflectance Pulse Oximetry for Blood Oxygen Saturation Measurement from Diverse Locations-A Preliminary Analysis". In: *MeMeA 2018 - 2018 IEEE International Symposium on Medical Measurements and Applications, Proceedings*. IEEE, June 2018, pp. 1–6. ISBN: 9781538633915. DOI: 10.1109/MeMeA.2018.8438781.
- [62] C. H. Shen, W. L. Chen, and J. J. Wu. "Research on multiple spectral ranges with deep learning for spo2 measurement". In: *Sensors* 22.1 (Jan. 2022), p. 328. ISSN: 14248220. DOI: 10.3390/s22010328.
- [63] S. J. Barker et al. "Measurement of carboxyhemoglobin and methemoglobin by pulse oximetry: A human volunteer study". In: *Anesthesiology* 105.5 (Nov. 2006), pp. 892–897. ISSN: 00033022. DOI: 10.1097/00000542-200611000-00008.
- [64] *Rad-57 Pulse CO-Oximeter*. Accessed: 10-02-2022. URL: <https://www.masimo.com/products/continuous/rad57/>.
- [65] S. Rhee, B. H. Yang, and H. H. Asada. "Artifact-resistant power-efficient design of finger-ring plethysmographic sensors". In: *IEEE Transactions on Biomedical Engineering* 48.7 (July 2001), pp. 795–805. ISSN: 00189294. DOI: 10.1109/10.930904.
- [66] H. Lee, H. Ko, and J. Lee. "Reflectance pulse oximetry: Practical issues and limitations". In: *ICT Express* 2.4 (Dec. 2016), pp. 195–198. ISSN: 24059595. DOI: 10.1016/j.icte.2016.10.004.
- [67] *BORA Band by Biosency*. Accessed: 10-02-2022. URL: <https://biosency.com/en/solution/>.
- [68] N. H. Lovell et al. "Biosignal quality detection: An essential feature for unsupervised telehealth applications". In: *12th IEEE International Conference on e-Health Networking, Application and Services, Healthcom 2010*. IEEE, July 2010, pp. 81–85. ISBN: 9781424463749. DOI: 10.1109/HEALTH.2010.5556528.
- [69] W. Karlen et al. "Multiparameter respiratory rate estimation from the photoplethysmogram". In: *IEEE Transactions on Biomedical Engineering* 60.7 (July 2013), pp. 1946–1953. ISSN: 00189294. DOI: 10.1109/TBME.2013.2246160.
- [70] M. A. Pimentel et al. "Toward a robust estimation of respiratory rate from pulse oximeters". In: *IEEE Transactions on Biomedical Engineering* 64.8 (Aug. 2017), pp. 1914–1923. ISSN: 15582531. DOI: 10.1109/TBME.2016.2613124.
- [71] W. Karlen, J. M. Ansermino, and G. Dumont. "Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. IEEE, Aug. 2012, pp. 3131–3134. ISBN: 9781424441198. DOI: 10.1109/EMBC.2012.6346628.

-
- [72] N. Pradhan, S. Rajan, and A. Adler. "Evaluation of the signal quality of wrist-based photoplethysmography". In: *Physiological Measurement* 40.6 (July 2019), p. 065008. ISSN: 13616579. DOI: 10.1088/1361-6579/ab225a.
 - [73] S. Vadrevu and M. Sabarimalai Manikandan. "Real-Time PPG Signal Quality Assessment System for Improving Battery Life and False Alarms". In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 66.11 (Nov. 2019), pp. 1910–1914. ISSN: 15583791. DOI: 10.1109/TCSII.2019.2891636.
 - [74] T. Pereira et al. "Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation". In: *Physiological Measurement* 40.12 (Dec. 2019), p. 125002. ISSN: 13616579. DOI: 10.1088/1361-6579/ab5b84.
 - [75] T. Pereira et al. "A Supervised Approach to Robust Photoplethysmography Quality Assessment". In: *IEEE Journal of Biomedical and Health Informatics* 24.3 (Mar. 2020), pp. 649–657. ISSN: 21682208. DOI: 10.1109/JBHI.2019.2909065.
 - [76] M. Elgendi. "Optimal signal quality index for photoplethysmogram signals". In: *Bioengineering* 3.4 (Sept. 2016), p. 21. ISSN: 23065354. DOI: 10.3390/bioengineering3040021.
 - [77] R. Krishnan, B. Natarajan, and S. Warren. "Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data". In: *IEEE Transactions on Biomedical Engineering* 57.8 (Aug. 2010), pp. 1867–1876. ISSN: 00189294. DOI: 10.1109/TBME.2009.2039568.
 - [78] A. Luke, S. Shaji, and K. A. Unnikrishna Menon. "Motion Artifact Removal and Feature Extraction from PPG Signals Using Efficient Signal Processing Algorithms". In: *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*. IEEE, Sept. 2018, pp. 624–630. ISBN: 9781538653142. DOI: 10.1109/ICACCI.2018.8554599.
 - [79] V. Periyasamy, M. Pramanik, and P. K. Ghosh. "Review on Heart-Rate Estimation from Photoplethysmography and Accelerometer Signals during Physical Exercise". In: *Journal of the Indian Institute of Science* 97.3 (Sept. 2017), pp. 313–324. ISSN: 09704140. DOI: 10.1007/s41745-017-0037-1.
 - [80] F. P. Lo and M. Q. Meng. "Double sensor complementary placement method to reduce motion artifacts in PPG using fast independent component analysis". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol. 2016-October. IEEE, Aug. 2016, pp. 3179–3182. ISBN: 9781457702204. DOI: 10.1109/EMBC.2016.7591404.
 - [81] H. Han and J. Kim. "Artifacts in wearable photoplethysmographs during daily life motions and their reduction with least mean square based active noise cancellation method". In: *Computers in Biology and Medicine* 42.4 (Apr. 2012), pp. 387–393. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2011.12.005.

- [82] Z. Guo et al. "A supervised machine learning semantic segmentation approach for detecting artifacts in plethysmography signals from wearables". In: *Physiological Measurement* 42.12 (Dec. 2021), p. 125003. ISSN: 13616579. DOI: 10.1088/1361-6579/ac3b3d.
- [83] G. Chan et al. "Multi-site photoplethysmography technology for blood pressure assessment: Challenges and recommendations". In: *Journal of Clinical Medicine* 8.11 (Nov. 2019), p. 1827. ISSN: 20770383. DOI: 10.3390/jcm8111827.
- [84] S. S. Mousavi et al. "Blood pressure estimation from appropriate and inappropriate PPG signals using A whole-based method". In: *Biomedical Signal Processing and Control* 47 (Jan. 2019), pp. 196–206. ISSN: 17468108. DOI: 10.1016/j.bspc.2018.08.022.
- [85] C. El-Hajj and P. A. Kyriacou. "A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure". In: *Biomedical Signal Processing and Control* 58 (Apr. 2020), p. 101870. ISSN: 17468108. DOI: 10.1016/j.bspc.2020.101870.
- [86] A. Visvanathan, A. Sinha, and A. Pal. "Estimation of blood pressure levels from reflective photoplethysmograph using smart phones". In: *13th IEEE International Conference on BioInformatics and BioEngineering, IEEE BIBE 2013*. IEEE, Nov. 2013, pp. 1–5. ISBN: 9781479931637. DOI: 10.1109/BIBE.2013.6701568.
- [87] E. Monte-Moreno. "Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques". In: *Artificial Intelligence in Medicine* 53.2 (Oct. 2011), pp. 127–138. ISSN: 09333657. DOI: 10.1016/j.artmed.2011.05.001.
- [88] J. C. Ruiz-Rodríguez et al. "Innovative continuous non-invasive cuffless blood pressure monitoring based on photoplethysmography technology". In: *Intensive Care Medicine* 39.9 (Sept. 2013), pp. 1618–1625. ISSN: 03424642. DOI: 10.1007/s00134-013-2964-2.
- [89] A. Suzuki and K. Ryu. "Feature selection method for estimating systolic blood pressure using the taguchi method". In: *IEEE Transactions on Industrial Informatics* 10.2 (May 2014), pp. 1077–1085. ISSN: 15513203. DOI: 10.1109/TII.2013.2288498.
- [90] S. Venkat et al. "Machine Learning based SpO2 Computation Using Reflectance Pulse Oximetry". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2019* (July 2019), pp. 482–485. ISSN: 1557170X. DOI: 10.1109/EMBC.2019.8856434.

-
- [91] S. Vijayarangan et al. "Robust Modelling of Reflectance Pulse Oximetry for SpO₂ Estimation". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol. 2020-July. IEEE, July 2020, pp. 374–377. ISBN: 9781728119908. DOI: 10.1109/EMBC44109.2020.9176410. arXiv: 2004.06301.
 - [92] H. Dubey, N. Constant, and K. Mankodiya. "RESPIRE: A Spectral Kurtosis-Based Method to Extract Respiration Rate from Wearable PPG Signals". In: *Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017*. IEEE, July 2017, pp. 84–89. ISBN: 9781509047215. DOI: 10.1109/CHASE.2017.64.
 - [93] W. Karlen et al. "Respiratory rate assessment from photoplethysmographic imaging". In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*. IEEE, Aug. 2014, pp. 5397–5400. ISBN: 9781424479290. DOI: 10.1109/EMBC.2014.6944846.
 - [94] P. S. Addison et al. "Pulse oximetry-derived respiratory rate in general care floor patients". In: *Journal of Clinical Monitoring and Computing* 29.1 (Feb. 2015), pp. 113–120. ISSN: 15732614. DOI: 10.1007/s10877-014-9575-5.
 - [95] S. Chugh and J. Kaur. "Low cost calibration free Pulse oximeter". In: *12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015*. IEEE, Dec. 2016, pp. 1–5. ISBN: 9781467373999. DOI: 10.1109/INDICON.2015.7443576.
 - [96] N. B. Hampson. "Carboxyhemoglobin: A primer for clinicians". In: *Undersea and Hyperbaric Medicine* 45.2 (2018), pp. 165–171. ISSN: 10662936. DOI: 10.22462/03.04.2018.3.
 - [97] M. A. Cretikos et al. "Respiratory rate: The neglected vital sign". In: *Medical Journal of Australia* 188.11 (June 2008), pp. 657–659. ISSN: 13265377. DOI: 10.5694/j.1326-5377.2008.tb01825.x.
 - [98] Z. Zhang, Z. Pi, and B. Liu. "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise". In: *IEEE Transactions on Biomedical Engineering* 62.2 (Feb. 2015), pp. 522–531. ISSN: 15582531. DOI: 10.1109/TBME.2014.2359372. arXiv: 1409.5181.
 - [99] *PLUX - Website*. Accessed: 05-02-2022. URL: <https://www.pluxbiosignals.com>.
 - [100] *Typing.com*. Accessed: 30-05-2022. URL: <https://www.typing.com/>.
 - [101] *Wim Hof Method - Website*. Accessed: 27-07-2022. URL: <https://www.wimhofmethod.com/breathing-exercises>.
 - [102] T. Citherlet et al. "Acute Effects of the Wim Hof Breathing Method on Repeated Sprint Ability: A Pilot Study". In: *Frontiers in Sports and Active Living* 3 (Aug. 2021). ISSN: 26249367. DOI: 10.3389/fspor.2021.700757.

- [103] Contec CMS50D Pulse Oximeter - Website. Accessed: 27-07-2022. URL: <https://contecmed.com/productinfo/602627.html>.
- [104] P. Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.
- [105] J. Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.
- [106] W. C. Navidi. *Statistics for engineers and scientists*. 5th ed. 2020.
- [107] F. J. Gravetter and L. B. Wallnau. *Statistics for the behavioral sciences*. 10th ed. 2015.
- [108] C. Carreiras et al. *BioSPPy: Biosignal Processing in Python*. 2015–. URL: <https://github.com/PIA-Group/BioSPPy/>.
- [109] I. I. Christov. "Real time electrocardiogram QRS detection using combined adaptive threshold". In: *BioMedical Engineering OnLine* 3.1 (2004). DOI: 10.1186/1475-925x-3-28.
- [110] M. Elgendi et al. "Systolic Peak Detection in Acceleration Photoplethysmograms Measured from Emergency Responders in Tropical Conditions". In: *PLoS ONE* 8.10 (2013), e76585. DOI: 10.1371/journal.pone.0076585.
- [111] A. Naguszewski and J. Weremczuk. "Determination of Oxygen Saturation Based on Distorted Photoplethysmography Signals by Scaling Method". In: *2018 15th International Scientific Conference on Optoelectronic and Electronic Sensors, COE 2018*. IEEE, June 2018, pp. 1–4. ISBN: 9781538641071. DOI: 10.1109/COE.2018.8435183.
- [112] D. G. Altman and J. M. Bland. "Measurement in Medicine: The Analysis of Method Comparison Studies". In: *The Statistician* 32.3 (1983), p. 307. DOI: 10.2307/2987937.
- [113] *Physical Activity Monitoring for Heart Rate (ANSI/CTA-2065)*. Consumer Technology Association, Feb. 2018.
- [114] D. Giavarina. "Understanding Bland Altman analysis". In: *Biochemia Medica* 25.2 (2015), pp. 141–151. ISSN: 13300962. DOI: 10.11613/BM.2015.015.
- [115] S. Moscato et al. "[In review] Photoplethysmography quality assessment for reliable heart rate estimate and morphological analysis". In: *IEEE Journal of Biomedical and Health Informatics* 22.15 (Aug. 2022), p. 5831. ISSN: 1424-8220. DOI: 10.3390/s22155831.
- [116] E. Britannica. *Occam's razor*. Accessed: 30-08-2022. URL: <https://www.britannica.com/topic/Occams-razor>.

- [117] D.-G. Jang et al. "A Simple and Robust Method for Determining the Quality of Cardiovascular Signals Using the Signal Similarity". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 478–481. DOI: 10.1109/EMBC.2018.8512341.

A

Informed Consent for Data Acquisition



NOVA SCHOOL OF
SCIENCE & TECHNOLOGY



Consentimento de Participação

Nº: _____

No âmbito da tese de Mestrado em Engenharia Biomédica, pelo mestrando Pedro de Bastos Veiga da *NOVA School of Science and Technology* (NOVA SST) em parceria com a *PLUX Wireless Biosignals*, coloca-se à consideração de V/Exa. a possibilidade de participação num estudo intitulado *“Multi-SpO₂ sensor hardware and embedded machine learning algorithm development for robustness improvement in SpO₂ measurements in wearables.”*

A oximetria de pulso é um monitor *standard* para inúmeras situações a nível clínico. Devido à sua simplicidade, é um dispositivo com um custo de produção baixo e de fácil utilização. Dado isto, uma aplicação natural deste tipo de equipamentos é em dispositivos *wearables*. No entanto, este mercado foca a utilização deste tipo de monitores apenas para a monitorização do ritmo cardíaco e ainda falta robustez nas medições da saturação de oxigénio no sangue (SpO₂) para aplicações clínicas. O objetivo desta tese de mestrado será o estudo do efeito de várias situações na qualidade do sinal obtido e em possíveis formas de o melhorar.

O presente estudo consista na obtenção de sinais ao longo de diversas atividades do dia a dia, como escrever no computador, escrever à mão e andar. Os sinais fisiológicos obtidos serão: **eletrocardiograma** (ECG), **respiração** (RIP), **saturação de oxigénio no sangue** (SpO₂) e **aceleração**. Estes sinais serão medidos usando os equipamentos da *biosignalsplux*. Estes equipamentos estão atualmente no mercado e cumprem todas as normas para dispositivos médicos. Irá também se proceder ao preenchimento de um questionário, onde serão registadas várias **informações pessoais** como a idade, sexo, padecimento de alguma patologia, entre outras informações.

A participação é livre, voluntária e sem custos, sendo que ao participante é facultada a possibilidade de desistir ou interromper a qualquer momento do estudo sem qualquer efeito negativo.

Nestes termos, é garantido o anonimato, assim como a confidencialidades dos resultados obtidos. Os resultados deste estudo serão mais tarde apresentados, no entanto nunca serão identificados os participantes.

Assinatura do Participante: _____

Data: ____/____/____





NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS

NOVA SCHOOL OF
BUSINESS