Evaluating Garmin Smartwatches Ability to Detect Oxygen Desaturation Events

Francesca Akpene Lumor

Thesis submitted for the degree of
Master in Programming and System Architecture
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2022

Evaluating Garmin Smartwatches Ability to Detect Oxygen Desaturation Events

Francesca Akpene Lumor

Evaluating Garmin Smartwatches Ability to Detect Oxygen Desaturation Events

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

Obstructive Sleep Apnea (OSA) is a common yet severely under-diagnosed sleep disorder that causes disrupted or reduced breathing during sleep. The gold standard and traditional way of diagnosing the disorder is with a polysomnography. This sleep study is however expensive and resource heavy, as it requires the patient to sleep in a laboratory with different physiological sensors attached to the body. Additionally, a sleep expert has to be present during the study and later scores the collected data. The CESAR project aims to reduce the time to diagnosis by creating an initial OSA detection at home. As smartwatches that are packed with sensors are becoming ubiquitous and connected to smartphones with large processing power, there are new opportunities emerging in the medical field with consumer electronics.

Pulse oximeter is one of many sensors used in polysomnography, and is now also located in various smartwatches. The sensor estimates blood oxygen saturation levels (SpO_2) which is relevant to OSA as reduced or disrupted breathing is associated with lowered oxygen levels. In this thesis, we use the Venu 2S Garmin smartwatch as test oximeter and the medical grade sleep monitor Nox T3 as reference oximeter. We test the usability and how Bluetooth connection losses are handled by the app for pairing and recording data from Garmin sensors. We collect data in non-invasive lab tests and overnight monitoring, and analyse how four Machine Learning (ML) classifiers (signal counting, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Random Forest (RF)) perform based on the metrics Cohen's Kappa (κ), accuracy, sensitivity and specificity. Lastly, we assess Garmin's sensor quality by the industry standard metric accuracy (A_{rms}), we calculate mean absolute error (MAE), and perform a Bland-Altman analysis to get the mean bias (mean of the difference) and precision (standard deviation of the difference).

The usability tests of the app resulted in descriptions like "easy to use", and we uncovered some ambiguities which were later fixed. As for the connection test, we were not able to reproduce the delay tolerance observed in previous tests. Classification based on signal counting is outdated and the achieved κ was the lowest of the classifiers (0.0). The best performance was achieved with RF with κ ranging between 0.35 - 0.57 on Nox data. All ML classifiers performed worse on Garmin data which could be attributed to the lower data quality. The accuracy of Garmin's sensor ranged from 1.305 to 9.883 which is a large variation (Mean = 3.01 $\pm$1.6). The signal quality did not meet ISO standard requirement of $\leq$ 3%, though not far

off. It did, however, reflect the typical A_{rms} specification under normal conditions described by the FDA. For individual recordings nearly 70% had accuracy lower than 3%. After grouping the data on different variables, we saw that the overnight data had the worst mean metrics but least variability, darker skin type had the worst mean and most variability, and movement and desaturation events were not correlated with worse accuracy.

Our classification results were affected by small samples. In the future, more data should be collected, preferably from overnight monitoring, in order to see more conclusive results. Despite rejecting all of our hypotheses of variables affecting signal accuracy, we observed some trends that should be looked into further.

Acknowledgements

First and foremost I would like to thank my supervisor, Professor Thomas Peter Plagemann. Our weekly meetings helped me structure the work and kept me motivated when it felt like it was getting overwhelming. Additionally, I would like to thank the CESAR project for allowing me to contribute to the important work that they have done, and are continuing to do, for Obstructive Sleep Apnea detection.

I would also like to thank every single person who volunteered for the studies. This includes those who actually performed the tests, and those who were just interested. This thesis would not exist without you, and your interest fueled me when motivation was getting low. Also, a special thanks to Garmin for providing the smartwatches and access to the Garmin Health Companion SDK.

Most importantly I would like to thank my family. My sisters, Josephine and Evelyn, for your support and encouragement throughout. An extra thanks also to Josephine for proofreading. The last and biggest thanks goes without a question to my mom, Margaret: **Akpe na Mawu be ènye danye!**

Acronyms

A_{rms} Accuracy root mean square. 6, 16, 21, 23, 24, 33, 43, 46, 48, 49, 58, 84, 103, 109

$COHb$ Carboxyhemoglobin. 14

CO_2 Carbon dioxide. 16

Hb Hemoglobin (deoxygenated or reduced). 14

HbO_2 Oxygenated hemoglobin. 14

Hz Hertz. 33, 34, 41

$MetHb$ Methemoglobin. 14

N_2 Nitrogen. 16

O_2 Oxygen. 14, 16

SaO_2 Arterial oxygen saturation. 13–17, 85

SpO_2 Peripheral oxygen saturation. 14, 16, 17, 21, 22, 32, 33, 43, 47, 58, 67, 74, 83, 84, 87, 88, 96, 97, 100, 102, 108

κ Cohen's Kappa. 47, 48, 60, 61, 74, 75, 78–83, 100, 101, 104, 109

AASM American Academy of Sleep Medicine. 9, 12, 17

AHI Apnea-Hypopnea Index. 9, 12, 13, 17, 29

AI Artificial Intelligence. 18

ANOVA Analysis of Variance. 51, 92–94, 102

API Application Programming Interface. 20, 30

CI Confidence Interval. 50

CSV Comma Separated Values. 29, 30, 32, 34, 57, 58, 60, 61, 64

CV Cross-validation. 45, 62, 74, 80, 83, 101, 104, 108, 109

DL Deep Learning. 18

ECG Electrocardiogram. 10, 13

EEG Electroencephalogram. 10, 13

EMG Electromyogram. 10, 13

EOG Electrooculogram. 10, 13

FDA Food & Drug Administration. 16, 17

FN False Negative. 24, 47

FP False Positive. 24, 25, 47

KNN K-Nearest Neighbours. 5, 45, 61, 74–76, 79, 81, 82, 104, 108, 109

LoA Limits of Agreement. 49, 50, 53, 62, 84, 85, 109

MAE Mean absolute error. 6, 21, 49, 62, 84, 89, 93, 103, 109

ML Machine Learning. 4, 5, 9, 18, 19, 21, 22, 41, 45, 46, 57, 60, 61, 74, 75, 79, 80, 82, 83, 100, 101, 108, 109

MVP Minimum Viable Product. 40, 99, 108

NN Neural Network. 18, 22, 41, 45, 82, 101

ODI Oxygen Desaturation Index. 12, 17, 29, 45, 59–61, 74–76, 81–83, 100, 108, 109, 111, 113

OSA Obstructive Sleep Apnea. 3–6, 9–11, 13, 17, 24, 43

PSG Polysomnography. 3, 4, 10, 11, 13

RDI Respiratory Disturbance Index. 12, 13

RERA respiratory effort related arousal per hour of sleep. 12

RF Random Forest. 6, 45, 46, 61, 74–76, 79–83, 108, 109

RIP Respiratory Inductance Plethysmography. 23, 27, 28, 42, 94

RMSE Root mean square error. 49, 103

SA Sleep Apnea. 6, 12, 18, 19, 21, 22, 30, 41–43, 45, 59, 61, 75, 104, 108, 113

SD Standard Deviation. 50, 84–86, 92, 93, 97, 110

SDK Software Development Kit. 20, 30, 31, 42, 55, 56, 100, 107, 108, 114

SVM Support Vector Machine. 5, 45, 46, 61, 74–76, 79, 82, 104, 108

TN True Negative. 24, 47, 76

TP True Positive. 24, 47, 76

UI User Interface. 56

UiO University of Oslo. 3, 6, 42

Contents

List of Figures

List of Tables

Listings

Part I

Introduction and Background

Chapter 1

Introduction and Motivation

1.1 Background and Motivation

All living beings sleep, and sleep has been shown to be essential for our cognition, physical and mental health [13, 59]. Not getting sufficient sleep can adversely effect our health, possibly leading to metabolic disorders, cardiovascular diseases and even occupational accidents [12]. While going to bed earlier and setting proper bedtime routines could help some people improve their sleep quality and quantity, for others the problem can be more serious than that.

Obstructive Sleep Apnea (OSA) is a condition where breathing is reduced or stopped involuntarily on multiple occasions throughout the night [66]. This is the most common form of sleep apnea where the airway has become narrowed, blocked, or floppy. These apnea episodes are often brief and therefore tend to be unknown to the person experiencing them, which makes it harder to detect and diagnose. Because of this, it is estimated that 70-80% of those affected remain undiagnosed [51].

The gold standard for diagnosing OSA and other sleep disorders is polysomnography (PSG). It utilises many different sensors for measuring physiological data during sleep, such as electroencephalogram, electrocardiogram, airflow, and oximeter to name a few. Though the sleep study is effective there are the downsides of it being quite expensive and requiring medical staff during the study and for analysing the data afterwards. These factors contribute to the high threshold for administering the study, which is also reflected in the number of undiagnosed cases.

Wearable technology has become widespread in recent years, with the most common type being smartwatches. These devices are packed with sensors that monitor the person wearing it and their surroundings. Some of them even have a pulse oximeter for monitoring the person's blood oxygen saturation. As earlier mentioned, people with OSA experience brief reduction or a complete stop in their breathing. This would lead to reduced oxygen saturation in the blood, which is why an oximeter is one of the sensors used in PSG.

The CESAR project is an interdisciplinary research project (computer science and medicine) at the University of Oslo (UiO) that aims to enable

monitoring of OSA at home for everybody [10]. The purpose of the project is to use low-cost consumer sensors for monitoring sleep, and in turn use smartphones for analyzing the collected data. Machine Learning (ML) will be used to detect desaturation events in the oximeter signals. The resulting analysis and scoring of events will give the person an indication of whether they should contact a physician. Furthermore, the analysis should give the physician a better foundation to decide whether a PSG should be performed. This process will hopefully reduce the time until diagnosis.

As previously mentioned, OSA is a severely under-diagnosed sleep disorder caused in part by an expensive and resource demanding procedure for diagnosis. The emergence of consumer electronics like smartwatches that are equipped with low-cost oximeters open a new market of possibilities in healthcare monitoring [17]. This is however not to replace the traditional PSG for diagnosis, but rather to allow for initial at-home detection of the disorder. In a previous thesis written by Felix Griffin Halvorsen, the use of Garmin smartwatches for initial detection was tested [26]. The results were promising, though the sample size was too small to draw any real conclusions. Another study by Lauterbach et al. [38] evaluating the quality of Garmin smartwatch's oximeter also found it promising, even claiming it to be a viable method to monitor blood oxygen saturation. This is however just the start, and there is still much to be researched on this topic.

With this thesis we aim to continue on Halvorsen's work by improving the algorithm for detection of desaturation events, and further assess the quality of Garmin's pulse oximeter. Other work in the CESAR project has used ML for classification [35, 36], and we will also do the same. Furthermore, an Android app, *Nidra* [60], has been previously created in the CESAR project for users to record, share, and analyze breathing data. If Garmin's sensor proves to be successful, then support for Garmin could be integrated into *Nidra*. In the end, we hope to bring the CESAR project a step closer on deciding whether using Garmin smartwatches in early detection of OSA is useful.

1.2 Problem Statement

Halvorsen's project [26] researched how to use Garmin smartwatches to detect desaturation events as part of OSA screening at home. Most of the focus was on implementing and testing an app for collecting data from Garmin sensors. The data collected was also assessed for quality and ability to detect desaturation events. There were some open problems left after his project. For instance, a script developed by Halvorsen to detect desaturation events needed to be reevaluated in order to be more effective. Another approach that was not explored was using more modern methods like ML for classification. Our thesis will therefore start where his left off and explore it further. The overall problem statement can be summed up in this one question:

- Can a Garmin smartwatch's pulse oximeter be used for initial at-

home sleep apnea detection?

The scope of this question is quite broad. We therefore break down the statement into answering these four questions:

1. Can the results from Halvorsen's experiments be reproduced?
2. Can using ML for classification improve event detection in Garmin signals?
3. Are the Garmin oximeters quality good enough for at-home OSA detection?
4. What external factors affects the quality of Garmin's pulse oximeter and how?

From these four questions we address three aspects related to the initial problem statement. The first is repeating relevant tests to see if the results can be reproduced. Since the app is the first entry into acquiring the signal data, we need to make sure it still functions as intended and is user-friendly. Volunteers were recruited through fellow students, project participants and others at the department through e-mails. They then performed a breathing experiment in the lab at the Department, and some also performed an unattended overnight sleep monitoring. After we acquire the data, we need to assess whether it can be used for prediction of desaturation events and OSA. The second question addresses this as the ability to correctly classify events in the recorded signal data. Questions three and four are related to the quality of the Garmin pulse oximeter. Kristiansen et al. [35] showed that the quality of the signals is important when it comes to classifier performance. We therefore try to evaluate different factors that could negatively impact signal quality.

1.3 Approach and Scope

Based on the outlined problem statement the main focus of this thesis is evaluating Garmin watches ability to detect OSA. The way this will be done is by first testing the app for usability and connection loss detection. As there is no previously existing large database of this nature the next step, most importantly, is acquiring data that can be analysed. We use the sleep monitor Nox T3 from Nox Medical [62] as the reference pulse oximeter. As a test oximeter, we use the Garmin smartwatch Venu 2S [20]. The signals of interest are oxygen saturation and accelerometer for synchronization. We collect data from two different methods, one of a more experimental nature in a lab, and the other from unattended overnight monitoring.

After acquiring the data we measure ability to detect desaturation events by using classifiers. Nox Medical has an accompanying software, Noxturnal [46], which automatically scores events. We use these labels as reference, and assess how the script created by Halvorsen and three ML classifiers (K-Nearest Neighbours (KNN), Support Vector Machine (SVM),

Random Forest (RF)) predict events. Since we collect signals from two devices, we also test the classifiers on them separately. This allows us to compare the performance on two different data sets.

Lastly, we assess the quality of the Garmin pulse oximeter. The metrics we assess the quality on is accuracy calculated by A_{rms}, which is the industry standard for pulse oximeters. We compare this with mean absolute error (MAE) as it is more robust when it comes to outliers. Then, we evaluate agreement between devices by performing a Bland-Altman analysis where we get the mean bias (mean of the difference) and precision (standard deviation of the difference). Furthermore, we try to assess if there are any external factors that significantly influence the accuracy. This will be determined through hypothesis testing.

A big challenge with such a project is finding enough participants. With the research being conducted during a pandemic it proposes a bigger challenge in this area. It could make finding a large and diverse group of participants more difficult. Furthermore, we do not have access to actual SA sufferers, which would be the most representative. Another restriction is the fact that the evaluation license we got from Garmin only allows us to use subjects from our organization, i.e University of Oslo (UiO). This means that it is probably not a representative sample. Preferably, more data would be collected from unattended sleep monitoring, as it is the most comparable to the actual use case. Because of time restrictions and limited access to subjects, we will also collect data from simulated lab tests.

1.4 Thesis Outline

This thesis is divided into five parts consisting of 11 chapters and appendices. An overview of the remainder of this thesis is presented in the following:

I Introduction and Background

- **Chapter 2 - Background:**

  This chapter gives a presentation of OSA; what it is, characteristics, prevalence, diagnosis, etc., and also presents the theory of essential concepts relevant in this thesis, i.e. on oximeters, desaturation events and ML.

- **Chapter 3 - Related Work:**

  There has been some research done on this topic before. Both within the CESAR project and otherwise. The methods and results of some relevant works are presented in this chapter.

- **Chapter 4 - Data Acquisition and Processing:**

  We present in this chapter the equipment that was used, the data we acquire from them and how they are processed.

II Method and Implementation

- **Chapter 5 - Method:**

  In this chapter, we present detailed information about the methods that we use. That is, the tests which are usability test, connection loss tests, classification testing and signal quality testing. Lastly we present the methods for analysis.

- **Chapter 6 - Implementation:**

  This chapter presents the implementation of the changes made to the app, and the Python scripts used for data processing, classification and metrics calculations.

III Evaluation

- **Chapter 7 - Results:**

  In this chapter, the results of the several tests performed will be presented.

- **Chapter 8 - Discussion:**

  In this chapter, we discuss and try to make sense of the results from the previous chapter.

IV Conclusion

- **Chapter 9 - Summary of Contributions:**

  In this chapter, a summary of our findings and contributions will be presented. We also address the problems stated in the introduction.

- **Chapter 10 - Open Problems:**

  In this chapter, we discuss open problems.

- **Chapter 11 - Future Work:**

  In this last chapter, we present future work that can be done.

V Appendices

- **Appendix A - Source Code:**

  A link to where the code and data sets that has been used in the thesis can be found.

- **Appendix B - Consent Agreement:**

  The consent agreement that had to be signed by the subjects before any data could be collected.

- **Appendix C - Usability Test Guide:**

  The guide which was followed in conducting the usability testing of the app.

- **Appendix D - Experiment Results:**

  All the results from both the sensor quality experiments and the classification experiments. The metrics are displayed for each recording and also mean for each category.

Chapter 2

Background

This chapter presents an overview of the different topics relevant for the thesis. We begin in Section 2.1 with describing what OSA is and how it is diagnosed. The next section, Section 2.2 goes in depth on the pulse oximeter, how it is relevant to OSA detection, different sensor types, and their quality. Section 2.3 presents the definition of desaturation events according to the American Academy of Sleep Medicine (AASM) and how this can be used to detect OSA. Next, we dive into ML, what it is and how it can be used to detect desaturation events in time-series data in Section 2.4. Lastly, we discuss how low-cost oximeter sensors, particularly those found in smartwatches, ties into this in today's society in Section 2.5.

2.1 Obstructive Sleep Apnea (OSA)

Sleep apnea is a sleep disorder marked by abnormal breathing during sleep [66]. The most common type is Obstructive Sleep Apnea (OSA) where the upper airway has become narrowed, blocked, or floppy. The illustration in Figure 2.1 shows how the airway can get blocked from the tongue and soft palate sliding back. This in turn leads to disruptive breathing, which can be described as being either reduced or shallow (hypopnea), or stopped (apnea). These involuntary lapses in breathing are called apnea events and occur on multiple occasions throughout the night. Additionally, after an apnea or hypopnea event, the body's oxygen level is lowered, also referred to as a desaturation event. Oftentimes, the person experiencing these apnea events awakens after around 10 seconds in order to breathe.

A study performed in Norway found that the prevalence of OSA is around 16% for Apnea-Hypopnea Index (AHI – measurement score for the severity of sleep apnea, defined as the number of apneas plus hypopneas per hour of sleep [23]) ≥ 5 and 8% for AHI ≥ 15 [29], which is consistent with the general population [23]. However, the sleep disorder often remains undiagnosed or is diagnosed late, with an estimation of 70-80% of those affected remaining undiagnosed. The high rate of undiagnosed cases can be attributed to the apnea events being brief, and thus the person experiencing them not remembering the awakenings the next day. Furthermore, a common symptom of OSA is feeling tired during the day,

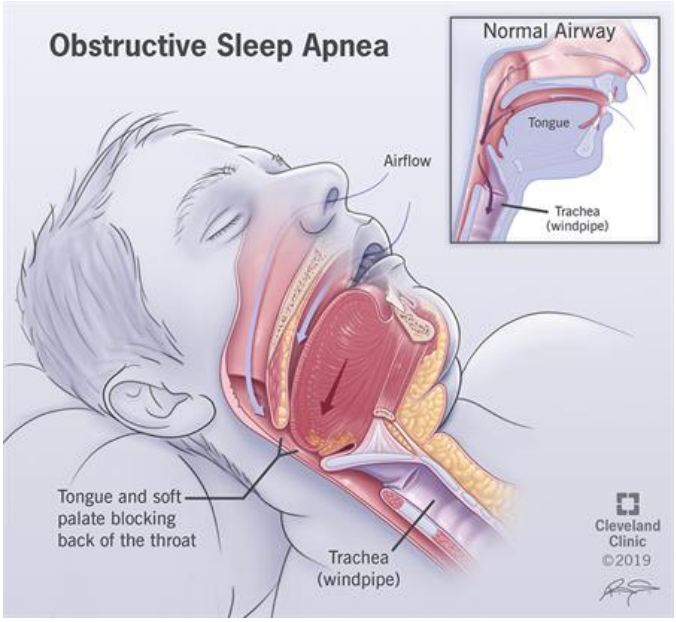


Figure 2.1: *Illustration of a blocked airway caused by OSA [47]*

which is fairly typical for many people.

Although the person suffering from OSA might not be aware of this, the consequences of the condition can be far-reaching. The disorder reduces the sleep quality through continuous interruptions throughout the night. Some common symptoms that arise from this include excessive daytime sleepiness, loud snoring, and morning headaches to name a few. On severe cases OSA has been linked to many different health problems such as sleep deprivation, depression, work accidents, and increased risk of cardiovascular and metabolic disorders [51, 66]. It is therefore evident that remaining undiagnosed can be detrimental to a person's mental and physical health.

2.1.1 Diagnosis

The gold standard for OSA diagnosis is the traditional laboratory study known as polysomnography (PSG). During the study, the patient sleeps overnight at a laboratory while various physiological parameters are being recorded. A sleep technician will also be present monitoring the study. An illustration of a PSG setup is given in Figure 2.2. PSG utilizes signals such as electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), and pulse oximetry, as well as airflow and respiratory effort [55]. With the corresponding sensors, PSG is able to record data such as sleep stages, limb movements, airflow, respiratory effort, heart rate and rhythm, oxygen saturation, and body position.

Also in Figure 2.2 at the bottom (B), we see an example of the generated polysomnogram from the sleep study. This polysomnogram reveals a desaturation event in the first row immediately after a long breathing

event in the second row. An analysis of the generated data is performed by medical personnel. The results could reveal any underlying sleep disorders, such as OSA, nocturnal seizures, narcolepsy, and periodic limb movement disorder to name a few.

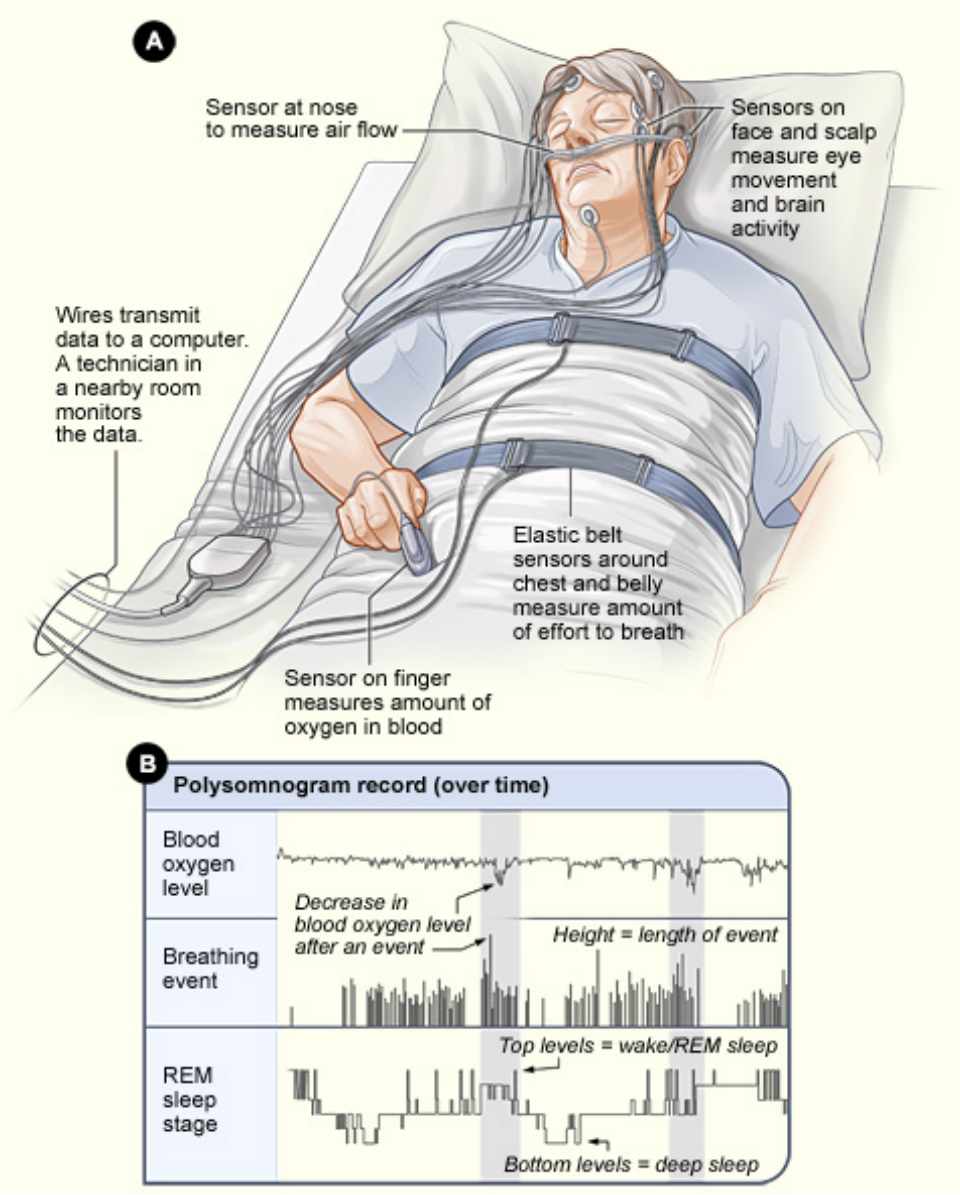


Figure 2.2: *Illustration of a PSG (A) and accompanying polysomnogram (B) [63]*

Despite PSG being the gold standard in sleep analysis, there are some downsides to the sleep study. For instance, they are quite expensive and require medical personnel for overnight observation and for later manual analysis of the data. Furthermore, even though the study is non-invasive, it can be quite uncomfortable for the patient. They would have to spend the night in an unfamiliar environment while being monitored by various machines and medical personnel.

Severity Ratings

The American Academy of Sleep Medicine (AASM) has published a set of rules for scoring respiratory events such as apnea and hypopnea in a Manual by Berry et al. [5]. Both events are required to last at least 10 seconds, while an apnea is accompanied with a $\geq$ 90% drop in signal from a respiratory sensor, and hypopnea with a $\geq$ 30% drop and $\geq$ 3% drop in oxygen saturation. Furthermore, the AASM specify metrics for determining the severity of SA. A common measurement is the Apnea-Hypopnea Index (AHI) which is the number of apnea and hypopnea events per hour of sleep. The metric further separates severity into these four classes:

- None/Minimal: AHI < 5 per hour
- Mild: AHI $\geq$ 5, but < 15 per hour
- Moderate: AHI $\geq$ 15, but < 30 per hour
- Severe: AHI $\geq$ 30 per hour

This is, by consensus, summed with the respiratory effort related arousal per hour of sleep (RERA) index to form the Respiratory Disturbance Index (RDI) metric given in Equation 2.1. Another measurement is through the Oxygen Desaturation Index (ODI) given in Equation 2.2, which is the number of $\geq$ 3% arterial oxygen desaturations per hour of sleep. These metrics (RERA, RDI, ODI) are optional.

$$RDI = AHI + RERA\ index \tag{2.1}$$

$$ODI = \geq 3\%\ arterial\ oxygen\ desaturations/hour \tag{2.2}$$

Types of Sleep Monitors

There are different types of sleep monitors that exist which are certified for medical use. Based on the signals that they collect and how they are used they have been classified as Type I, II, III, or IV [71]. What is required to be classified within each type varies between different definitions. The first classification system was given by the AASM in 1994. However, it is now outdated given the widespread of home sleep testing devices. We will in the following present the definition of the type given by Center for Medicare & Medicaid Services (CMS):

- **Type I**

  A Type I monitor requires the sleep study to be performed in the laboratory. A sleep technologist needs to be present overseeing the

study. This is the category PSG falls under, and all the previously mentioned signals required for PSG must be included along with additional channels.

- **Type II**

  For Type II monitoring devices, a full PSG can be performed outside of the laboratory, like at home. The main difference from PSG is that a technologist does not need to be present. A minimum of seven signals are required which are EEG, EOG, ECG/heart rate, EMG, airflow, respiratory effort, and oxygen saturation.

- **Type III**

  Type III monitors can also be used at home unattended. The four signals that must be included are two for respiratory effort/airflow, ECG/heart rate and oxygen saturation.

- **Type IV**

  Type IV monitors can also be used unattended at home. They must include a minimum of three signals that allow direct calculation of an AHI or RDI score through measurements of airflow or thoracoabdominal movement. If other information is used to derive AHI or RDI, then it must be approved.

There exists many gadgets with sensors such as pulse oximeters accessible to consumers which can be used during sleep to monitor signals. These gadgets are not included in the classification system as they are not certified for medical use.

At-home Sleep Apnea Detection

From the classification of sleep monitors, it is evident that there are many portable alternatives to PSG that allow for OSA detection at home. One such device is the Nox T3 sleep monitor (Type III) [69] that has been used previously in the CESAR project as the reference device [19, 26, 36, 40]. Such devices are quite expensive and are not considered as consumer devices for the average person. In a review by Mendonça et al. [43] they found that commercial devices can be used as an initial OSA diagnosis tool. They should, however, not replace PSG for patients with high comorbid medical conditions or sleep disorders, or patients over 65 years old. Additionally, there are limitations surrounding at home devices' ability to detect OSA during REM sleep. They have a tendency to under-diagnose cases of mild OSA and they have a higher failure rate compared to PSG [43, 55]. Such devices should therefore not be used in place of PSG for OSA diagnosis, but rather for initial detection.

2.2 Pulse Oximetry

As earlier mentioned, one of the many physiological signals monitored during PSG is arterial oxygen saturation (SaO_2). This is obtained through

pulse oximetry. Pulse oximetry is a non-invasive method to estimate SaO_2 by reading the peripheral oxygen saturation (SpO_2). Both red and infrared light is emitted through human tissue, and the amount of light absorbed in the blood results in an estimate of oxygen (O_2) in the blood. The method usually measures the values through sensors attached to a finger or earlobe. This will be elaborated further in the following section, which will be based on the work of Wukitsch et al. [79], Tremper and Barker [70], and Nitzan, Romem and Koppel [45].

Firstly, to better understand pulse oximetry, the technology behind estimating SaO_2 will be explained in more detail. Spectral analysis is the ability to detect elemental composition by defining the unique light absorption. In other words, we can detect different elements by their absorption of light. The method is based on the Beer-Lambert law that states that the concentration of absorbant in solution can be determined as a mathematical function of the amount of light transmitted through the solution. When it comes to measuring SaO_2, it is possible because the protein hemoglobin (Hb) which is found in the red blood cells, transports oxygen (O_2) to body parts that need it. Oxygenated hemoglobin (HbO_2) has a different absorption, which is referred to as extinction, than deoxygenated hemoglobin (Hb or reduced Hb). This can be seen in their extinction curve at red (wavelength 650 to 750 nm) and infrared (wavelength 900 to 1000 nm) light in Figure 2.3. At the red wavelengths HbO_2 has less extinction than Hb, and the reverse is true at a lesser extent for infrared wavelengths. To summarize, we estimate SaO_2 as the ratio of the difference in extinction of red an infrared light between HbO_2 and Hb. The formula for calculating SaO_2 is given in Equation 2.3.

$$SaO_2 = \frac{HbO_2}{HbO_2 + Hb} \tag{2.3}$$

As pulse oximetry is a non-invasive method, it is important to understand that the light absorption is affected by skin, other tissue, and venous blood. With this knowledge in hand, it is possible to determine SpO_2 from the level of absorption at these wavelengths, and isolating the signals coming from arterial blood from other tissues. Light absorption in the different tissues are constant over time (direct current (DC)), while in pulsating arterial blood it is variable (alternating current (AC)). What we are interested in measuring (SaO_2), is in the AC. We therefore need to scale divide the AC level by the DC level to get the corrected AC value. This whole process is what is referred to as pulse oximetry.

Pulse oximeters are calibrated against a CO-oximeter, the gold standard for measuring SaO_2. In Figure 2.3, we see two other "types" of hemoglobin, carboxyhemoglobin ($COHb$) and methemoglobin ($MetHb$), which are not measured by a pulse oximeter. The CO-oximeter, in addition to measuring Hb and HbO_2, can also measure $COHb$ and $MetHb$. It differs from the pulse oximeter by being invasive, as the CO-oximeter measures the

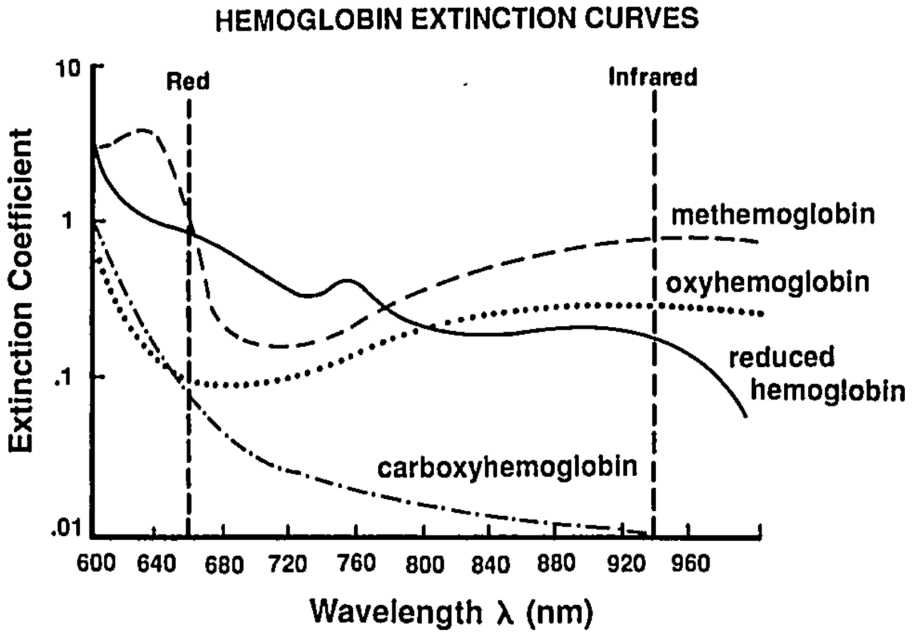


Figure 2.3: *Hemoglobin extinction curves of Hb, HbO_2, COHb and MetHb at different light absorption [70]*

aforementioned values through the absorption of light passed through a blood sample. Being able to measure all these values along with direct access to the blood accounts for why it gives a more accurate measure of SaO_2.

2.2.1 Measurement

A pulse oximeter consists of two light sources, red and infrared, and a photodiode, which is a light detector. Different types of pulse oximeters fall into two different groups, which are transmittance and reflectance. In the transmittance group we have devices where the two light sources are on opposite side of a photodiode with the measurement site between them. Examples of this are the previously mentioned finger probe or ear probe. With reflectance oximetry, both light sources and the photodiode are on the same side, and light is reflected back to the photodiode through the measurement site. A smartwatch with a pulse oximeter would fall into this group.

The transmissive method for pulse oximetry is the most commonly used, with the finger being the most used placement. In [45], they observed conflicting results in different studies of which performed better between reflectance oximetry with measurement site on the forehead and transmittance oximetry. Many smartwatches are adding pulse oximeters to their list of many sensors. According to a study performed by Lee et al. [39] the reflective mode with the wrist as measurement site does not perform as well as the finger probe, and the wrist is not an optimal measurement

site. There is, despite this, promising research on the quality of Garmin smartwatch's oximeter with one study claiming it to be a viable method to monitor blood oxygen saturation [38].

2.2.2 Estimating Oximeter Quality

The accuracy of pulse oximeters is determined by the difference between the SpO_2 measured by the pulse oximeter and the SaO_2 measured by a CO-oximeter. The standard metric used for calculating accuracy for pulse oximeters is the accuracy root mean square (A_{rms}) difference as specified by the ISO 80601-2-61:2017 [30]. The terms A_{rms} and accuracy are used interchangeably for the same thing in regards to oximeter quality assessment. The accepted accuracy for pulse oximeters in the range 70-100% as specified by the ISO standard is $\leq 3\%$. Most manufacturers specify an accuracy of about 2%. In a study by Milner and Mathews [44], of 758 sensors in use in 29 NHS hospitals in the UK, 169 (22.3%) had inaccuracies $> 4\%$.

The Food & Drug Administration (FDA) has in a guidance document given recommendations on how to test oximeter accuracy *in vivo* under laboratory conditions [72]. There should be an even spread of 200 or more samples between 70-100% SaO_2 from a pulse oximeter and CO-oximeter. This should be taken from a minimum of 10 healthy subjects varying in age and gender, where a minimum of 15% has dark skin pigmentation. They also recommend the use of Bland-Altman plots for visualising agreement for individual subjects and all subjects. The most common way of achieving an even spread of SaO_2 samples is through breathing in a gas mix containing nitrogen (N_2), carbon dioxide (CO_2) and O_2. Varying the concentration of oxygen and nitrogen in the gas mix leads to more stable plateaus of SaO_2. The ISO also suggests the use of non-invasive testing against another pulse oximeter that is traceable to a CO-oximeter.

The FDA recognizes that the accuracy of a pulse oximeter is affected by external factors such as patient characteristics, application site, and sensor geometry [72]. They have therefore outlined typical accuracy that can be expected between measured values (SpO_2) and reference values (SaO_2) for SpO_2 ranging between 70-100%. This is listed in Table 2.1. From the expected accuracies, we see that transmittance oximetry actually gives the best estimate of SaO_2.

Sensor Type	A_{rms}
Transmittance, wrap and clip	$\leq 3.0\%$
Ear clip	$\leq 3.5\%$
Reflectance	$\leq 3.5\%$

Table 2.1: *Typical A_{rms} Specification by Sensor Type [72]*

2.2.3 Limitations

There are many limitations when it comes to estimating SaO_2 through pulse oximetry, some of which have been highlighted by the FDA [50]. This spans from the inner implementation of a specific oximeter, to environmental factors in the particular setting it's being used in. We will mostly focus on those related to environment and human physiology.

Skin pigmentation has been mentioned to affect pulse oximeters ability to measure SpO_2 [79]. There has been some research on exactly how the effect results in reality [6, 61]. Cold temperatures of the measuring site also affects SaO_2 readings, also nail polish should not be worn for a finger probe reading. Another problem with pulse oximeters is that movement artifacts can reduce the reliability. There are some challenges also when it comes to measuring internal measures from a wristwatch. For instance, how tight or loose the watch is worn, or a change of position may affect the sensor's performance [39]. The watch moving on the wrist is to be expected during sleep. There is also the thickness of the skin on the wrist and the fit of the watch.

2.3 Desaturation events

In a study performed by Gries and Brooks [25], they found that the average oxygen saturation during sleep for healthy patients tends to be around 96%. Patients with OSA significantly differed from this with a mean SpO_2 at 93.5%, and also registered values as low as 65.9%. This decrease in blood oxygen level is a direct cause of breathing events. Oxygen saturation above 94% during sleep is considered normal [73]. There are no generally accepted classification of oxygen desaturation severity, though below 90% is considered mild ranging to severe below 80% [74].

2.3.1 Desaturation Recommendations From AASM

As previously mentioned, AASM has published a set of rules for scoring respiratory events such as apnea and hypopnea [5]. These rules also include what classifies as a desaturation event. For an event to be classified as a desaturation, there are three criteria that need to be met:

1. The event has to last a minimum of 10 seconds,
2. 3% SpO_2 decrease compared to baseline, and
3. baseline is defined as the mean of stable breathing in the two minutes before event onset.
    (a) In the absence of stable breathing, we use the three largest breaths in the preceding two minutes.

In a review by Rashid et al. [54] they investigated the use of ODI for OSA diagnosis compared to AHI. Despite the many differences

in measurement definitions between the studies, they concluded that consideration should be given for diagnosing adult OSA with a 4% ODI of ≥ 15 events/hour and for recommending further evaluation for diagnosing OSA with a 4% ODI ≥ 10 events/hour.

2.4 Machine Learning (ML)

ML is a subfield of Artificial Intelligence (AI) (Figure 2.4) focused on the use of various self-learning algorithms that uses knowledge from data to predict outcomes. The process often consists of a set of data called *training data*, of which we extract features from. We then use the features to train a *model* to classify the data based on those features. Afterwards, the trained model can be used on new data. The model consists of a specified algorithm that, in the case of classification, identifies what class the data is based on the features. The outcome of the model is which class the data belongs to. A real-world example which ML is often used for is spam filtering. ML approaches have traditionally been categorised as *supervised learning* (labeled), *unsupervised learning* (unlabeled) or *reinforcement learning* (reward-based) based on the use case and available training data.

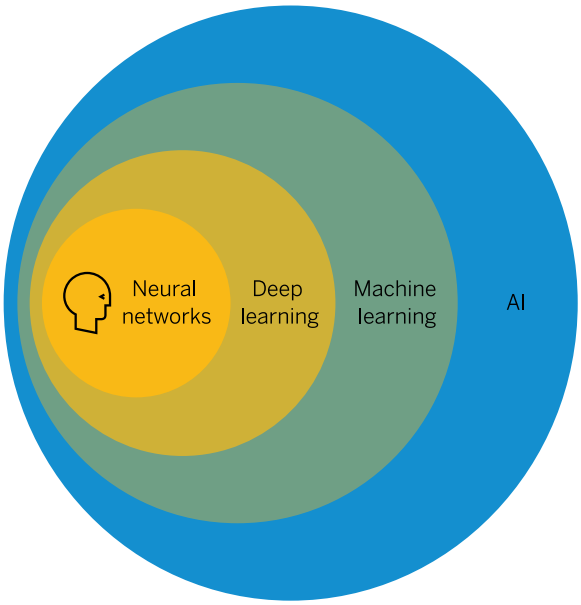


Figure 2.4: *Diagram of AI and related subfields ML, DL, and NN [75]*

A subset of ML is Deep Learning (DL) and Neural Networks (NN), which is a form of ML that is inspired by the the human brain structure. It differs from ML as it does not require human intervention for, e.g. feature extraction and labeling, and consists of multiple layers, from input to output and hidden layers in between. There has been much use of DL algorithms in SA detection, and they significantly outperform traditional ML classifiers. As these will not be used in this thesis because of the small volume of data, we will not describe the methods any further.

2.4.1 Supervised Learning

As previously mentioned, supervised learning is one of the main types of ML models. It uses labeled data sets for training the algorithms that will later be used for classification. This means that the class we want to give the data is known before training the model. The algorithms and methods we use in this thesis fall into this category.

2.4.2 Time Series Classification

A *time series* is a sequence of values or data points over a period of time. The data points are often an observation of a feature or event that varies over time, and the collected data points usually have evenly spaced intervals. An example of this is the temperature of the weather in a period of a year, sampled every day. Usually, ML with time series data is related to identifying trends in the data, forecasting future values, or predicting the class of a new time series based on the class labels of a set of time series [2]. Another use of ML with time series is for labeling segments of the data which is called *period-based* classification. This is a more fine-grained approach where the individual data set is divided into periods with class labels. In regards to SA detection, period-based classification can be used to label oxygen saturation as either apneic or normal breathing [36]. One thesis further investigated an even finer-grained technique known as *segmentation* for labeling events [32].

2.5 Low-Cost Consumer Sensors

Oximeters are available for consumers at a fair price. They are however not a common device to have. This is not the case for smartwatches, which are fast becoming ubiquitous. As this thesis is focusing on smartwatches, the low-cost sensors addressed in this section will be the ones found in smartwatches.

A smartwatch or fitness tracker is a type of non-invasive wearable technology. As other wearable devices, they are powered by microprocessors and can be described as minicomputers that can be worn on the body [9]. Wearable devices are still fairly new in techgear, however the market for smartwatches has grown large the last years. In 2019, the market was valued at $20.64 billion (approx. 171.29 billion NOK) and will continue to grow in the future, projected to reach $96.31 billion by 2027 (approx. 799.29 billion NOK) [67]. Some leading smartwatch brands include Apple, Samsung, Fitbit, and also Garmin [64, 67].

A commonality with smartwatches is that they are packed with an abundance of sensors. For instance, the *Vivoactive 4* Garmin smartwatch contains an accelerometer, light sensor, gyroscope, heart rate monitor, GPS, thermometer, and even a pulse oximeter to name a few [21]. With these sensors, the watch can monitor different aspects of health such as sleep, stress level and heart rate. It can even alert the wearer if the values get concerning [21]. These sensors and features make smartwatches a

good companion for tracking fitness progress and monitoring changes in personal health.

Smartwatches are inexpensive compared to medical grade devices and are more accessible and accepted among consumers. Some of them also allow for third parties to develop applications for them by providing a REST Application Programming Interface (API) and Software Development Kit (SDK). Furthermore, smartwatches are often paired with smartphones through WiFi or Bluetooth which allows for offloading heavier computation and data processing to the smartphones. This is also the case with smartwatches automatically syncing to the cloud.

There are some challenges and limitations associated with using wearable technology in healthcare monitoring, some of which have been discussed in the article by Dunn et al. [17]. There is the fact that the battery life tends to be short, which in this case, could hinder the smartwatches' ability to monitor throughout the night. Ensuring the user's data privacy and security is maintained is important because of the sensitive nature of the health data being collected. Furthermore, for these devices to be beneficial to the user in pre-clinical health monitoring then the sensors should be accurate.

Chapter 3

Related Work

There has been work done related to several topics we touch upon in this thesis. We highlight in this chapter some of the most relevant works. In Section 3.1, we cover three articles that research the quality of low-cost sensors, two of them specifically address watches. Section 3.2 presents two works on ML for SA detection. Lastly, Section 3.3 presents three previous theses written in the CESAR project which have directly influenced this thesis.

3.1 Low-Cost Sensor Quality

The quality of pulse oximeters is heavily researched, especially in relation to the Covid-19 pandemic [3]. We have chosen two works [27, 38] that are the most similar, and will also later be compared to our results. The third work presented is included as it gives some insight on quality of pulse oximeters with the wrist as measurement site.

Lauterbach et al. [38] researched how heart rate and SpO_2 measurements by Garmin Fenix 5X Plus watch perform and different simulated altitudes compared to the medical grade oximeter Nonin WristOx2. 23 subjects sat in a customized environmental chamber that simulated altitudes of 12,000; 10,000; 8,000; 6,000; and 900 ft. The mean bias calculated from a Bland-Altman analysis was at its worst at 3.3% for the highest altitude. For the lower altitudes the mean bias ranged between 0.7% to 0.8%. These results are promising for Garmin's pulse oximeter.

Harskamp et al. [27] tested the quality of ten popular low-cost pulse oximeters. SpO_2 signals from 35 patients were compared to blood samples. The pulse oximeters were evaluated based on the metrics A_{rms}, MAE and the Bland-Altman analysis with mean bias. None of the oximeters met the ISO-standard requirement of $\leq 3\%$ A_{rms} with the lowest being 3.9, and the mean bias ranged from -0.6 to -4.8. Despite this, they could accurately rule out hypoxaemia.

In the work by Phillips et al. [49] they explore the reliability of SpO_2 measurements from the wrist. This was done by creating a custom wrist-worn pulse oximeter and testing a baseline algorithm for SpO_2 calculation compared to a more advanced algorithm. Data was collected

from 10 subjects who wore the custom wrist-worn pulse oximeter on their dominant hand and a fingertip sensor on the other. They develop a solution called $WristO_2$ that selectively prunes unreliable data, as the existing algorithms used in fingertip SpO_2 sensors lead to over 90% of readings being inaccurate in wrist sensors. With the help of ML, the model was trained on the reliability label based on the agreement between the devices ($\pm$2.0 percentage points agreement is classified as reliable), and tested on the custom pulse oximeter. The final result was an order of magnitude reduction in error at the cost of less frequent readings compared to existing algorithms.

3.2 ML for SA Detection

A lot of work exists on ML for SA detection, however we will only highlight two that are directly relevant to our work. The first one by Kristiansen et al. [36], uses different combinations of 27 classifiers and four sleep signals to evaluate classification performance and resource use. The data used was from the A3 study which consists of more than 7,400 hours of data from unattended sleep monitoring at home by 579 patients. The classifiers were grouped into traditional, recurrent NN, and feed-forward NN and hyperparameters were varied for each classifier. The models were then tested on signals (nasal airflow (N), oxygen saturation (O), and respiratory effort from the chest (C), and abdomen (A)) that were grouped into 60-second periods. Each period was then labeled as either apneic or abnormal. The results in short was that only using oxygen saturation produced as good performance as all signals with 0.8543 accuracy (kappa: 0.7080). Also, the NN models outperformed the others irrespective of model architecture and size.

In another work by Kristiansen et al. [35], they investigate the classification performance of five ML algorithms (Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), and Random Forest (RF)) on sleep data (respiratory signals from the nose, abdomen, chest, and oxygen saturation). To study the impact of data quality on classification performance, they used both high-quality data from the Apnea-ECG database and low-quality data from the MIT-BIH database, both hosted by PhysioNet [37]. The results show that data quality significantly impact classification performance, as the accuracy for the MIT-BIH data was over 90% while for the MIT-BIH data it was in the range 60-70%. The low result for MIT-BIH was due to noise in the data, smaller sample size, and some class imbalance.

3.3 Previous CESAR Theses

Some of the previous theses written in the CESAR project have laid the foundation for what will be done in this thesis. Either directly or indirectly by their results or methods. This section will highlight three of them

by Fredrik Løberg [40], Kenneth Aune Frisvold [19] and Felix Griffin Halvorsen [26].

3.3.1 Measuring the Signal Quality of Respiratory Effort Sensors for Sleep Apnea Monitoring

Fredrik Løberg evaluated in his thesis the quality of two respiratory effort sensors [40]. The two sensors under test were piezo-electric effort belt (PZT) from BITalino and an impedance plethysmography (IP) sensor from Shimmer. These sensors were then compared to a reference which was Respiratory Inductance Plethysmography (RIP) sensor from Nox Medical. A signal capture procedure was created with the purpose of collecting data from subjects in a shorter amount of time compared to a full overnight monitoring.

The purpose of the signal capture procedure was to acquire data representing different forms of disrupted breathing from external subjects. Different requirements were set in order to best simulate a representative sleep monitoring session. For instance, the subject was required to be lying in a bed and sleep in a supine (back) and lateral decubitus (side) positions. When it came to the disrupted breathing, the subject was required to perform three different breathing styles in addition to breathing stops which represents apneic events. The styles were normal breathing, shallow breathing representing hypoapneic events and deep breathing. Each of these events (breathing stops, shallow and deep breathing) were to last between 10-20 seconds in order to be detected, while normal breathing should last longer than these. The signal capture procedure including all of these requirements lasted around 16 minutes.

The metrics used for quality evaluation are related to the ones used by medical personnel which are sensitivity, positive predictive value (PPV), and clean minute proportion (CMP) for breath detection accuracy, along with the breath amplitude accuracy metric weighted absolute percentage error (WAPE). The work in his thesis was later published as an article [41].

3.3.2 Non-Invasive Benchmarking of Pulse Oximeters - An Empirical Approach

There is an emergence of gadgets and sensors that are easily accessible to consumers which could be used for medical purposes. Determining the quality of these sensors can be expensive and require special expertise. For instance, pulse oximeters traditionally require subjects to breathe in a gas mix before the blood is analysed by a CO-oximeter. Kenneth Aune Frisvold set out to create a non-invasive benchmarking protocol for pulse oximeter evaluation [19]. The two low-cost pulse oximeters Cooking Hacks MySignals (CH) and BITalino were compared to the more expensive reference Nox T3 sleep monitor. The metrics assessed were the industry standard for accuracy A_{rms}, and precision and mean bias from a Bland-Altman analysis. Lastly, an apnea detection analysis was performed with Nox T3 as the reference.

Frisvold's benchmarking protocol required the subjects to follow a breathing script slightly different from Løberg's, which would result in eight simulated apneas. The experiment starts with three minutes of stabilizing calm breathing followed by breath holding for a minimum of 10 seconds then two minutes of calm breathing. The breath holding followed by two minutes calm breathing would be repeated eight times. The results reveal the importance of feedback to the subjects during the experiment to make them aware of the process. Also, training beforehand is vital for the success. The benchmarking protocol was performed by ten subjects and the accuracy of the CH pulse oximeter was 1.29%($\pm$0.8).

3.3.3 Garmin Smartwatches to Detect Desaturation Events as Part of OSA Screening at Home

Felix Griffin Halvorsen wrote in his thesis about Garmin smartwatches' ability to detect desaturation events [26]. The purpose was to investigate the potential of consumer smartwatches, such as Garmin, for OSA detection at home. In order to assess this it required implementing an app for signal data collection from a Garmin watch. Furthermore, data was collected from Garmin watches and the Nox T3 pulse oximeter as reference simultaneously and used for classifying events and evaluating signal quality. As our thesis can be seen as a continuation of that one, we will give a more detailed overview of it. A total of five different tests were performed: user testing, detecting connection loss, energy efficiency, sensor evaluation with a breathing script and an overnight test. Four Garmin smartwatches featuring pulse oximeters were tested, i.e., *Vivosmart 4*, *Vivoactive 4*, *Venu* and *Fenix 6 Pro*. After some preliminary tests of the watches, only Venu and Fenix 6 Pro were included in the oximetry experiments.

The user testing uncovered that the text might be too small for people with poor eyesight, some improvements were needed in the paired watches screen as well. Otherwise, the users were satisfied with the app. In testing the app's ability to reconnect after a connection loss, a delay tolerance was discovered, meaning some data was buffered. The energy efficiency test resulted in both the smartwatch and Garmin watches lasting at least throughout the night.

The smartwatches' pulse oximeter went under laboratory tests with the signal capture procedure created by Løberg and overnight sleep tests. In both tests the two smartwatches and a Nox T3 device were all worn at the same time. A total of eight individuals performed the lab test, while the overnight experiment only included one test subject. Values for sensitivity, accuracy and specificity were calculated for both tests. Each 30-second window was classified as True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN). The accuracy metric A_{rms} was used for signal quality assessment.

From the lab test the Nox T3 detected a total of 24 desaturations in all eight tests. The overnight sleep monitoring was performed by a person with no previous history of OSA. The test lasted 5 hours and 49 minutes,

and the Nox T3 detected three desaturation events during that time. There were artefacts introduced in the test, such as movement and light. The affected part of data was removed before resampling and interpolation.

The smartwatches accurately detected some desaturation events as could be seen from the signals. It occurred on some occasion that the script however, did not detect all the desaturation events in the experiment with the breathing script. The Venu watch had a more stable performance compared to Fenix, though Venu had a high rate of FPs in the overnight monitoring. The Fenix watch's performance on the eight tests can be summarised as inconsistent. The reason for this could be that the events were not long enough or severe enough to trigger an event as defined by the script. Furthermore, a falsely low oxygen saturation reading before a desaturation event reduces the chances of detecting the event. There were also many outliers and a number of false positives and false negatives.

Comparing the script to the Noxturnal software suggests that the software uses a less strict definition than the one Halvorsen's script is based on. If the readings had been stable, then the script would be able to detect all events. In reality, oxygen saturation fluctuates and the watches are not that stable. The overnight test best represents how the smartwatches will be used in real life. Unfortunately, this experiment was only performed once on one subject. From the other experiment it is evident that the results differ from test to test. Therefore, it is important to investigate further the watches' performance in overnight experiments with more test subjects. Furthermore, total number of subjects in the experiments was small, which makes it difficult to draw any conclusions.

Chapter 4

Data Acquisition and Processing

In this chapter, we present the equipment we use to collect data and accompanying software (Section 4.1), and also how we process and prepare data for analysis (Section 4.2).

4.1 Equipment and Software

4.1.1 Nox T3

Nox T3 is a portable medical grade sleep monitor produced by Nox Medical [69]. It is categorised as a Type III monitor which means it is certified for medical use unattended at home. The supported sensors include dual Respiratory Inductance Plethysmography (RIP) belts, ECG/heart rate, nasal air flow pressure, pulse oximeter, accelerometer, snore sensor, to name a few. The device is battery-powered with a single AA battery. It is widely used for diagnosing sleep disorders such as sleep apnea.

(a) *Nox T3* **(b)** *Nonin WristOx2*

Figure 4.1: *The monitor and wrist oximeter from Nox Medical [62]*

4.1.2 Nonin WristOx2

Nonin WristOx2 is the wireless Bluetooth connected pulse oximeter linked to Nox T3 [78]. This device is also battery-powered by two AAA batteries. The oximeter is worn on the wrist as seen in Figure 4.1b and a finger is inserted into the attached finger probe. It provides pulse rate and blood oxygen saturation signals which start recording immediately when a finger is inserted into the probe. The device features the Nonin PureSAT technology that removes undesirable signals and filters the signals in challenging conditions such as low perfusion filter or motion to improve measurement accuracy [52].

4.1.3 Wearing Nox Equipment

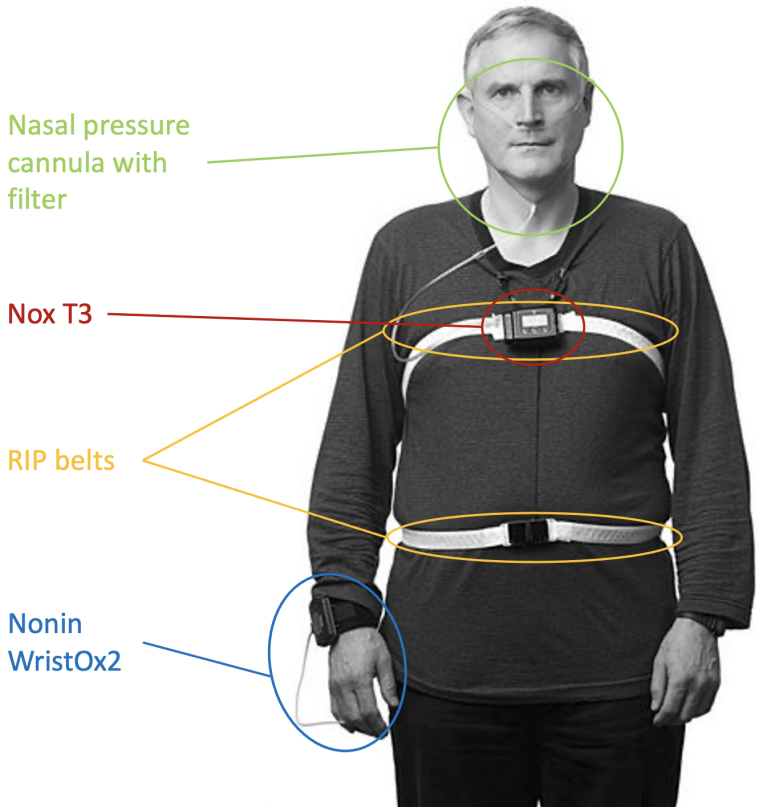


Figure 4.2: *Nox T3 - Fully worn equipment*

Figure 4.2 shows how the equipment is worn on the body. The main monitor is placed on the patient's chest with clips to fasten it in place. Attached to the monitor in the pressure connector is a nasal cannula which measures air flow pressure. The cannula is placed in the nose, fastened behind the ears of the patient, and tightened under the chin. The two RIP belts attached to the device are worn around the chest and waist. These accurately determine breathing effort. Lastly, the Nonin WristOx2 is worn on the wrist with an attached finger probe and connected to the monitor via Bluetooth.

This medical grade device is used as the gold standard in our studies. While the actual pulse oximeter is the Nonin WristOx2, we refer to the pulse oximeter and the collected data as Nox or Nox T3 data/signals. The signals we will be using include blood oxygen saturation and accelerometer.

4.1.4 Noxturnal

The Noxturnal software system is the associated system to Nox Medical devices [46]. Noxturnal allows a user to configure the device, download and analyse the data. This is done by connecting the device to a computer that has the software installed through a USB cable.

Before starting a new recording the Nox T3 device needs to be configured. During configuration a new recording session is set up with a patient profile where specifying either a name or a patient ID is required. This is also where the oximeter is linked to the monitor. After a recording is finished the data can then be uploaded to the profile on the software. The recorded signals are automatically extracted, processed, annotated and scored by the software's algorithms. An AHI and ODI score is automatically generated from the software's algorithms, as well as other sleep parameters like sleep stages, duration and efficiency. The data is visualized in reports and sheets as can be viewed in Figure 4.3. This allows for user-friendly and easy use for the clinicians. Furthermore, the clinicians can create their own sheets, add notes and manually analyse the data. Finally, the data can be exported to various formats such as excel or CSV.

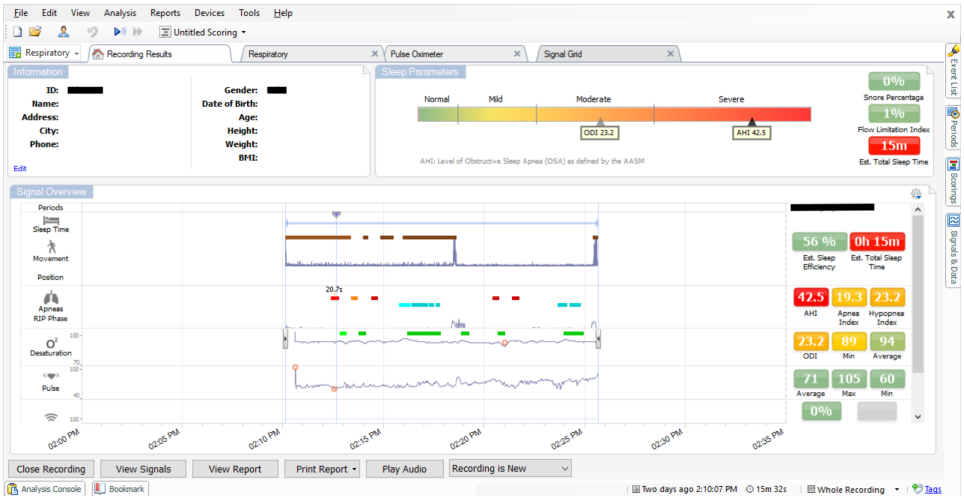


Figure 4.3: *Screenshot of recording in Noxturnal software*

A downside to the software is that it can only be installed on a Windows machine. Furthermore, the software is the only way to access the data from the devices. The device also needs to be configured between each recording if it is on the same day. Starting a new recording after one has ended appends the data to the same session. This makes it slightly more difficult to export the sessions as they will all be in the same file. It is therefore

not possible to perform multiple recordings on the same day for the same profile as separate sessions.

Data Acquisition

Since we want to compare the signals collected by Nox T3 to signals from other sensors, we need the data in another format than what is provided by the software. We therefore have to export the data in a compatible format. As previously mentioned the data undergoes some processing by the Noxturnal software. What this entails is unknown to us. As the software is the only access we have to the data, this means we don't have direct access to the raw data from the sensors. Despite this, we refer to the data from Noxturnal as the raw data.

We add the data we are interested in into a separate sheet. This includes a timestamp in the format of one second per row. Of sensor data we include accelerometer and oxygen saturation. These are in the format of the mean of that second timestamp. We also include the three events of desaturation, movement and artifact. The format of the events are binary, meaning for each row there is a 1 in the column if there is an event and 0 otherwise. The sheet is then exported as a CSV-file named as the ID of the recording.

The events are automatically scored by Noxturnal's algorithms. Usually a specialist will also manually score the data afterwards, but this was not done in our case. A study performed by Kristiansen et al. [34] concluded that the difference in AHI was small, and the automatic scoring classified sleep recordings with more than 90% accuracy into SA categories. Nox Medical themselves report about the great accuracy and reliability of their respiratory analysis compared to manual scoring in recent publications [46].

4.1.5 Garmin Venu 2S

Venu 2S is the second iteration of the Venu series from Garmin [20]. The S denotes that it is the smaller model in the series. The watch is a advanced companion for health and exercise monitoring as it has several sensors such as GPS, compass, thermometer, pulse oximeter, to name a few. Additionally it has different functions for specific activities such as running, golfing, biking, swimming, and allows for planning and analysing the activities. A list of some of the specifications can be viewed in Table 4.1 while a comprehensive list can be viewed on the Garmin website [20].

The data that has been collected from the sensors can be viewed on both the watch itself or on the Garmin Connect app. To access the data outside of these two platforms access to the Garmin Companion Software Development Kit (SDK) or Application Programming Interface (API) is necessary.

Features	Venu 2S
Weight	38.20g
Size	40mm
Connectivity	Bluetooth®, ANT+®, Wi-Fi®
Battery life	Smartwatch mode: 10 days Battery saver mode: 11 days GPS mode + music: 7 hrs GPS mode - music: 19 hrs
Accelerometer	Yes
Pulse oximeter	Yes (spot-check, *all day and sleep)
Internal memory	200 hrs of activity data

Table 4.1: *Garmin Venu 2S Specifications. * These are optional*

4.1.6 Garmin Health Companion Software Development Kit (SDK)

Garmin provides a Companion SDK which gives access to real time data from connected devices. The Companion SDK allows developers to create Android or iOS applications that can stream real time health data from the sensors. We get access to the SDK through Garmin's developer portal. The SDK provides an example app for initial testing and getting familiar with the setup and extensive documentation. A license key is required for the app to work. With the SDK, an app has been implemented with the purpose of collecting sensor data for later analysis.

4.1.7 Cesar smartwatches - Real Time App

The app Cesar smartwatches was implemented by Halvorsen [26] with the purpose of recording and storing the sensor data recorded from the paired watch. It is an Android app that has implemented the Garmin Health Companion SDK. A user of the app can pair their Garmin device and start and stop a recording session on the chosen device. The app consists of the four activities `MainActivity`, `PairingActivity`, `ShowPairedActivity` and `CollectActivity`, two classes and an interface which can be seen in Figure 4.4. We will give a brief presentation of the app and its implementation based on what each of the four activities do.

MainActivity

The `MainActivity` provides what is the start screen of the app. The screen only contains a button for further navigation. Other than that the activity checks for permissions that are required for the SDK.

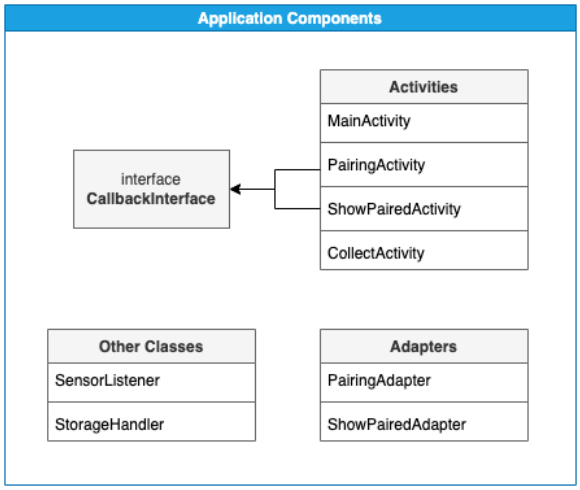


Figure 4.4: *App architecture overview*

PairingActivity

From the start screen we can navigate to `PairingActivity`. On this screen we see a list of unpaired Garmin devices that are nearby. The items on the list are clickable and clicking on one will pair the device to the app. The activity displays a list containing devices scanned from the area. The list, which is implemented as a `RecyclerView` gets populated with the help of an adapter. There is also a button on this screen that takes the user to a new activity with a list of paired devices.

ShowPairedActivity

The `ShowPairedActivity` screen looks mostly the same as the previous screen with a list of clickable device names. The difference is the listed devices are already paired. There is no separate button on this screen but clicking a list item will take you to the next and last activity `CollectActivity`.

CollectActivity

`CollectActivity` handles data collection. It starts with a start recording button on the middle of the screen. Clicking this button starts a new recording and a SensorListener class listens to the chosen sensors. At the same time the layout switches to a screen with a button to stop recording.

Data Acquisition

We collect all the raw data from the watches sensors. This includes SpO_2, accelerometer, heart rate, heart rate variability and respiration. The data are stored as individual CSV-files on the device. Of these we are interested in SpO_2 and accelerometer. The most important columns from the SpO_2

data set are the timestamp, duration and value which contains the signals at $1Hz$ sample rate.

4.2 Processing Data

To summarize, the collected data that we use from Nox is the SpO_2 signals, accelerometer signals, and events for desaturation, artifact and movement. From Garmin we use the SpO_2 signals and accelerometer signals. There are still some preprocessing that will need to be done with the data. For instance, the data should be adjusted to the same length, synchronized and preferably stored in the same file for easier use later on. What the preprocessing fully consists of will be presented here.

4.2.1 Data Formatting

Before any analysis of the data can be done there needs to be some formatting of the data. An obvious factor is removing outliers. For SpO_2 signals obvious outliers include values above 100% and below 0%. Additionally the signals have to be on the same frequency if they are not already to be comparable.

As we are comparing how two different sensors read internal values at the same time, it is required that the data sets have the same length. This circles back to synchronization as we are comparing how these methods read the same internal values. If one of the sensors have recorded more signals than the other, then those signals do not have any equivalent values to be compared to from the other sensor. Parts of the data sets that do not have a corresponding signal from the other sensor will therefore be removed. An important part in the preprocessing of the data is synchronizing the signals from the two devices. A more in depth description of this will be presented next.

4.2.2 Synchronization

As we are comparing signals from two different pulse oximeters there is a need for them to be synchronized first. An initial approach is to start the recording on both devices at the same time. They do however have different internal clocks so this is close to impossible as a synchronization method. Additionally, from looking at the data sets the Nonin oximeter and Venu 2S pulse oximeter tend to start picking up signals later than when the recording immediately starts.

The methods for synchronization are based on two methods proposed by Frisvold [19]. The devices were synchronized by aligning the peaks in the oximeter data. If there were no peaks in the oximeter data, then additional synchronization was done with the accelerometer. Both of these were compared in addition to visually inspecting the plotted oximeter graphs. Finally, we fine-tuned the shift in data that resulted in the best accuracy in A_{rms}.

4.2.3 Preprocessing Steps

To summarize in a broad outline, the preprocessing steps consists of:

1. Resampling the signals to $1Hz$,
2. interpolating missing data,
3. removing outliers,
4. synchronizing the signals,
5. combining columns and equalizing lengths, and
6. storing data in new CSV-file.

These steps need to be taken before any further analysis of the data can be performed.

Part II

Method and Implementation

Chapter 5

Method

Halvorsen performed several tests and experiments in his thesis. Some of these previous tests will be reevaluated to see if the results can be reproduced and if improvements could be made. In this chapter, we describe the methods that we use for testing and evaluation. In Section 5.1 we present the repeated app tests (usability and disconnection), how they were previously performed and improvements we have made. Section 5.2 addresses all the methods related to testing the Garmin signals. This includes collecting data, classification, and estimating quality. Lastly, in Section 5.3 we address the methods we use for analysing the data.

5.1 Application Testing

The app tests performed by Halvorsen include usability testing, connection loss detection between the smartwatch and the phone, and energy use testing of the phone. The purpose of these tests were to see if the app met the set requirements, both functional and non-functional. The tests to be repeated are usability testing of the real time app and Bluetooth connection loss detection. The goal of repeating these two tests is to see if we can reproduce the previous results and also get more insights. The main requirement for repeating these tests is to follow the same procedure. At the same time, there should be room to evaluate ways that these tests can be improved to produce new information. We list the following requirements that should be satisfied for conducting the app tests.

- **Equipment**

  For both tests it is required that the app, smartphone and smartwatch work as intended. This means the equipment has enough battery, the app is installed on the phone and it is possible to pair the phone and start/end a recording in the app.

- **Participants**

  A more comprehensive definition of usability and usability testing is given in Subsection 6.1.2. In broad strokes, we will test the app on the intended user to assess some predefined metrics. The participants

recruited for the usability test need to be from the intended user population. Furthermore, to ensure some representation we want participants with different levels of technology expertise.

- **Test instructor**

  For the usability test, we have a test instructor who will conduct the test. The role of the test instructor is to observe the "evaluator" when performing the tasks, assist when any problems arises and ask questions to better understand and get feedback.

- **Usability test process**

  To ensure consistency and comparability of the test with different participants, there needs to be a clear guide of the process. In this guide, the roles need to be clearly defined as well as the purpose of the test. Everyone should be aware of what data is being collected and how, and what is being tested which is the app and not their performance. Lastly, they should be aware of their right to withdraw whenever they want and have all data deleted.

5.1.1 Usability Testing

Usability is defined by ISO 9241-11:2018 as follows: *"the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* [31]. It is also further specified that usability is not a specific attribute of the product/system/service, but rather an outcome from interacting with it. Usability testing is defined by the ISTQB as testing to evaluate the degree to which the system, product or service meets the definition of usability by ISO 9241-11:2018 [24, 58]. In other words, usability testing evaluates a target audience's performance on a system in the specified context based on certain metrics. It essentially gives us an indication of the quality of the system or product based on user feedback.

The system that the target audience will be evaluated on is the real time app. It has been argued that mobile devices require some different components for usability testing. This was introduced in the PACMAD usability model by Harrison et al. [58]. The PACMAD model extends the ISO standard's three attributes (effectiveness, efficiency, satisfaction) [31] with learnability, memorability, errors and cognitive load.

Previous usability test

The focus for the previous usability test was on measuring different qualities of usability for a first-time user of the app. Following are the metrics and questions that were assessed in the previous usability test:

- **Efficiency** - time completing the task and on each screen.
- **Effectiveness** - number of clicks to solve the task.

- **Errors** - number of errors, experience any crash.
- **Satisfaction** - easy to use, responsive, satisfied with the experience, readability.
- **Learnability** - easy to learn, improvement of effectiveness and efficiency second time using.

Table 5.1 shows the results of Halvorsen's usability test. With this test the focus was mainly on giving the metrics a quantitative score. There were also some follow-up questions that allowed more elaboration from the subjects. For instance, one subject remarked that the small text in the app could possibly be difficult for people with bad eyesight to read. This has been improved ahead of more user testing.

Test Results				
Metric	**Subject A**	**Subject B**	**Subject C**	**Subject D**
Amount of Clicks	5	6	5	5
Time spent	8s	8s	15s	11
Time per screen	ĩs	ĩs	̃3s	̃2.2s
Errors occurred	0	1	0	0
Satisfaction	5	5	5	5
Ease of use	5	5	5	5
Responsiveness	5	4	5	5
Readability	4	5	5	5
Learnability	5	5	5	5

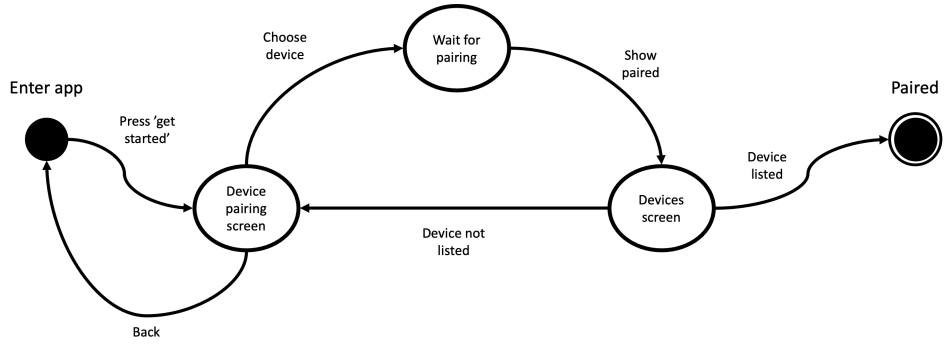
Table 5.1: *Usability results from Halvorsen [26]*

Current focus

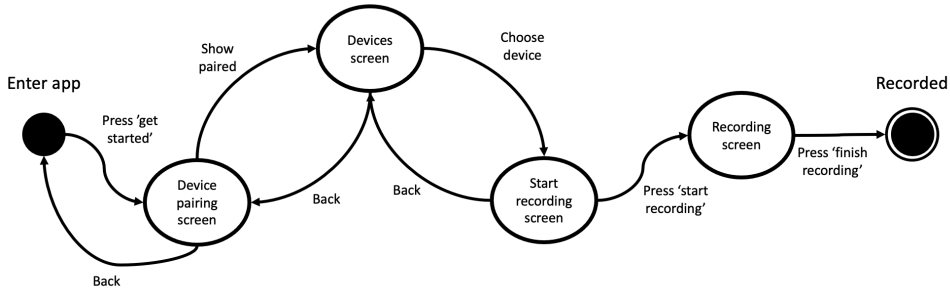
From the results of the previous usability test the app was proven to be fairly simple and straightforward. As the app has not significantly changed since the last round of tests, and the simple functionality of the app we decide to take a more qualitative approach. A more qualitative approach of the apps usability can give us different and valuable insight. For this round of usability testing the questions will be more focused around satisfaction.

There are mainly two different actions one can currently take in the app; pairing the device and recording data. Both of these actions are visualised in state-transition diagrams in Figure 5.1. The diagrams show that the device needs to be paired first before a recording can be performed. Therefore, the tasks will be performed in this order in the usability test.

The participants will complete the two tasks while outwardly describing their actions. After each task they will get some questions pertaining to



(a) *pairing a device*



(b) *performing a recording*

Figure 5.1: *State-transition diagrams of the two tasks to be completed*

their experience. Finally, they will get some time to talk about their opinion of the app in its entirety. The questions will essentially be focused on their impression of the app and if they see any potential improvements. The goal is to ask open-ended questions that allows the participant to explain their thoughts on the app. An in-depth guide can be found in Appendix C. It will be a moderated usability test allowing for interaction with the participant [68]. As the app is still a Minimum Viable Product (MVP) there are some known improvements that could be made. We are interested in uncovering some early issues that can improve the usability.

5.1.2 Connection Loss Detection

There are different scenarios in which the devices may disconnect during recording. For instance, the user could wake up in the middle of the night to get some water without bringing their phone with them. The range between the smartwatch and the phone could therefore exceed what is allowed for the Bluetooth connection. Another scenario is the body is blocking the signals between the watch and the phone while asleep. This could happen at certain angles of sleep and there could also be an added barrier of thick covers.

Halvorsen tested connection loss by distance with an Android smartphone and a Garmin watch. After recording had started the phone was left behind whilst the person wearing the watch went outside of communication range. According to Garmin support this range is within 10 meters of

the device without environmental factors such as walls blocking or interference from other electronic devices [76]. The test included two different lengths for the duration of disconnect; a short one of one minute and longer one of one hour. These will be replicated for this round of testing. Additionally, we will test if the 10 meter distance is enough for the devices to disconnect.

Another form of connection loss that will be tested is by blockage over a shorter distance. This will be a simulation of the signal blockage that could happen during sleep. The phone will be placed 1-2 meters away while the person wearing the watch will place their body and covers over it. The goal is to see if the watch disconnects from the phone with these barriers despite the close proximity. This will last for five minutes.

For each test the data will then be examined for any jumps in duration in the accelerometer data as it has the most stable sampling rate of $25Hz$.

5.2 Signal Testing

Halvorsen evaluated the signal quality of both experimental data and sleep data. He also evaluated the desaturation classification performance of a signal counting script. In addition to these previous tests that were performed, we have taken it a step further. For desaturation event classification we also include the use of ML algorithms. We will compare the performance of these algorithms to the original script, and also of the different ML algorithms used. In the assessment of sensor quality, we also test different hypotheses of variables that affect the measured Garmin signal accuracy.

We collect data, analyse the quality of said data and use it to detect events. How we define the process and the metrics we use to evaluate our results needs to be defined beforehand. There are some requirements set by the ISO 80601-2-61:2017 [30] regarding the sample and method for in vivo testing of accuracy of a pulse oximeter. These requirements, and also the FDA's guidance document [72], set the foundation for our data collection and quality assessment, although we are not following this to a tee. We present in the following some requirements and also limitations for collecting, using, and analysing the signal data.

- **Limitations and scope**

  As we are mostly collecting data from lab tests, and none of the data is from people with SA, it puts some limitations in regards to classification. The lab recordings are of a shorter length, which means even if we perform many tests it will not result in a large quantity compared to overnight monitoring. Because of this, we can not use algorithms such as NN, instead we will use traditional ML algorithms.

  When it comes to lab assessment of pulse oximeters, the ISO standard requires the subject to breath in a gas mix to simulate different oxygen saturation values. Because of lack of access, we use a breathing script

for subjects to simulate breathing events. Furthermore, it is required to be compared to 200 blood samples evenly from 70-100%. As we are using another pulse oximeter as reference, we will instead set a minimum data set size of 200 signals.

- **Privacy**

  The participants are required to sign a consent agreement before any data is collected. The consent agreement contains all information about the project, including what data is being collected from the subjects which is physiological data. The agreement can be found in Appendix B. Furthermore, they are informed of their freedom to withdraw any collected information at any time.

- **Equipment**

  For signal data collection we will be using two different sets of devices. With Nox T3 we need to have one AA battery for the monitor and two AAA batteries for the pulse oximeter. Also the monitor needs to be configured with a new profile before each recording. Before recording Garmin signals we need to ensure that the smartwatch and phone has enough battery, and that the app is working as intended. For hygienic reasons, the nasal cannula and RIP bands are switched for each subject, and all equipment are cleaned with disinfectant.

- **Test population**

  Because of restrictions put in place for using Garmin's smartwatch and accompanying SDK, our population will be people associated with UiO. Within this, we have set the requirement of 15% with a dark skin type as specified by the Fitzpatrick scale [28].

- **Information comprehension**

  There are a lot of things that can go wrong during lab recordings and overnight monitoring. We therefore emphasize the importance of information comprehension by everyone involved. This is especially important with unattended overnight monitoring where the subject will have to handle all equipment set-up by themselves.

- **Structured signal capture procedure**

  The process for the lab recording needs to be structured so that we can control certain aspects and replicate for several subjects. The subjects are required to lie down, replicating sleep. They should not be talking to avoid introducing unnecessary artifacts, and they should lie as still as possible unless otherwise specified. The procedure should be easy for the subjects to follow during the test. To ensure this, the next and/or current step should be made easily available to the subject. Also, we want for the lab recordings to be representative of SA while still being controlled, so they should simulate both apneic and hypopneic events.

- **Data characteristics**

  The signals from the devices need to be comparable. It is therefore important that they have the same format and that they are synchronized.

  In order to classify events in the SpO_2 signals, it is required that there is enough variation in the data to register as desaturation events. This depends entirely on the ability of the subject to simulate breathing events well enough to cause oxygen desaturation. Additionally, the overnight monitoring sessions will be performed by subjects that, to our knowledge beforehand, are healthy and without any signs of SA. We can therefore not predetermine the level of oxygen saturation in the range 70-100%, either from the lab tests or from overnight monitoring. What we can control is that the number of desat events is equal to the normal oxygen level events for classification. To ensure that there is also enough data to train and test the models on, it is required that each data set has a minimum of two desat events.

- **Reproducibility of classification performance**

  The results and the process of producing the results needs to be reproducible. Therefore, the classifiers and data need to be explained and accessible to others.

- **Quality metrics**

  Accuracy calculated in A_{rms} is the industry standard metric for pulse oximeters. Most pulse oximeters in use have an A_{rms} accuracy of 2%. The ISO standard states it should be $\leq$ 3% for SpO_2 values between 70-100%, while the FDA specifies that for reflectance oximeters it typically is $\leq$ 3.5%. We will use the $\leq$ 3% as the preferred accuracy limit and the $\leq$ 3.5% specified by the FDA as the upper limit of acceptable accuracy values.

5.2.1 Signal Capture Procedure

Fredrik Løberg originally created a signal capture procedure, also referred to as a breathing script, which has later been used in several CESAR theses [26, 40], also in different adaptations [19]. The purpose of the script was to simulate breathing events such as apnea and hypopnea that can occur during sleep, and the resulting desaturation event after such breathing events. We are mostly interested in desaturation events as we are only testing the pulse oximeter's oxygen saturation signals. However, since we are assessing the ability to detect OSA, we also require the data sets to be similar to a person who has sleep apnea.

The full duration of the procedure is 15 minutes. During these minutes the subject will lie first in a supine position and about halfway through turn to their side. Additionally, the subjects will have to perform three different breathing styles in addition to breathing normally. These are holding their breath, shallow or short breaths and deep breathing. Each of these will last for 17 seconds. The breathing script can be seen in Table 5.2.

Position	Time	Style
back	min 1-2	normal
	min 3	hold (17 sec)
	min 4	hold (17 sec)
	min 5	shallow (17 sec)
	min 6	deep (17 sec)
	min 7-8	normal
side	min 9-10	normal
	min 11	hold (17 sec)
	min 12	hold (17 sec)
	min 13	shallow (17 sec)
	min 14	deep (17 sec)
	min 15	normal

Table 5.2: *Breathing script/signal capture procedure*

Every subject is asked if they are able to hold their breath for 17 seconds before the procedure. Additionally, every session starts with a walkthrough of the procedure that will be shown to the subject as PowerPoint presentation during the procedure. Furthermore, during the walkthrough they are required to test shallow breathing and deep breathing to ensure their comprehension of the styles. The procedure is done in a laboratory at the University with the subject lying down in a sofa. The watch will be worn three different ways during the breathing script testing.

Lastly, there will be a test of the watch's performance in the intended natural setting. That is, participants will perform an unattended overnight sleep monitoring at home. The same equipment will be worn as for the breathing script. The difference is the participant will wear this overnight and go to sleep as they would usually do. The watch will be worn as the subject finds is most natural for themselves. This is the most realistic performance measure, and will give a more accurate view of the watch's performance. As the subjects will be putting on the equipment and starting the recording on both devices themselves, they will get a lesson on how to do so correctly before the monitoring. In case anything is forgotten, they are sent a video on Nox T3 hookup as a preventive measure. Afterwards, the subjects will report on any significant issues regarding the equipment or any other factors relevant to the sleep monitoring.

5.2.2 Desaturation Event Classification

In Halvorsen's thesis he tested a script that used a signal counting method to score desaturation events in data sets [26]. It is based on the ODI definition of desaturations, therefore this script is called the ODI classifier in this thesis. This method for classification is naive and outdated with the emergence of methods like ML and NN. We will therefore in this thesis include other methods for classification. As there is not enough data for extensive NN models, we have chosen to only use simple supervised ML algorithms. The algorithms that will be tested include K-Nearest Neighbours (KNN), Support Vector Machine (SVM) and Random Forest (RF).

The ODI classifier and ML classifiers performance will be compared to the analysis in Noxturnal. This will be done by running all the classifiers on the SpO_2 recorded by both Nox T3 and Venu and comparing the results with the automatic labelling in Noxturnal. We will from this be able to compare how these classification methods perform on data from a medical grade device and a consumer device.

For the ML classifiers we use a hold-out test set and 10-fold Cross-validation (CV) to estimate classification performance. For the hold-out test set, we test individual recordings while the rest of the data is used for training the model. We use CV as well, as it gives a better indication of how well the model performs as it performs hold-out testing ten times on different splits of the data set.

Despite having a breathing script where we aim to create events, there are no guarantees that it will result in any desaturations. As for overnight monitoring, the subjects included in this experiment had no prior SA diagnosis. In the case where there are no desaturation events, the most valuable information will be assessing the quality of the signals.

K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is the simplest of all ML classifiers [48]. The concept of the algorithm is easily understood by it's name; an object is classified based on the majority class of the k number of nearest neighbours. For instance, if we have a fruit and want to determine whether it is an apple or a pear, we can see which fruit the five nearest fruits placed on a two-dimensional feature space have. The main considerations of the algorithm is therefore distance and k. The k is the only parameter the algorithm takes and can be varied for best predictions. Having a small k can result in overfitting while having a larger k can lead to underfitting. Another modification is adding weights to the neighbors.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised ML algorithm that tries to find a hyperplane that best separates two classes [57]. It does this by finding the margin that results in maximum distance between the classes.

The algorithm uses only the points that are closest to the hyperplane, which are called support vectors. Continuing with the fruit example, the apples and pears are placed on a two-dimensional space based on their features. The SVM algorithm will try to find the best hyperplane that separates the apples and pears. We classify the fruit we have based on which side of the hyperplane it falls. If the classes are completely separated, then we can use a linear SVM. The more common is a non-linear SVM where we use kernel tricks to find the best fit to separate the classes. Some common kernels are the polynomial, sigmoid and RBF kernels.

Random Forest (RF)

The Random Forest (RF) classifier builds decision trees on different samples and the selected class is the class which is chosen by most trees [56]. That is, RF uses an ensemble technique that uses the decision trees algorithm to make predictions. Decision trees are also a supervised ML algorithm. It is a binary tree that starts from a root node and splits the data set recursively into nodes based on conditions of each decision node. We are finally left with leaf nodes that represents the classes. With RF we use a bagging method to split the original data set into subsets for each decision tree. To exemplify with the fruit again, we create decision trees with five different subsets of the original data. The trees choose which fruit it is based on feature conditions. If, for instance, four of the five trees classify the fruit as an apple, then the RF algorithm lands on the fruit being an apple.

5.2.3 Sensor Evaluation

As earlier mentioned, fitness and smartwatches are not intended for medical use. There is a high level of quality that is required of devices that are used for diagnosis. Garmin has written a disclaimer stating that their pulse oximeter is not intended for any medical purpose including diagnosing, treating, curing or preventing any disease or condition [1]. We will therefore evaluate how good the Garmin watch is compared to a Type III device. The standard metric that is used to evaluate oximeter quality according to ISO 80601-2-61:2017 [30] is accuracy calculated by the A_{rms} metric. We will calculate this along with several other metrics for each data set. Which metrics that are used and why can be found in Subsection 6.3.2.

There are some known variables that could affect the performance of a pulse oximeter, especially one with the measuring site on the wrist. For instance, the physical characteristics of the user, the fit of the device, or the placement of the watch on the wrist could affect the pulse oximeter's ability to read blood oxygen levels [50]. There are several different factors and we will be testing some of them. On account of this the watch will be tested with three different placements. The first being what the subject finds most comfortable, the second one is tight and the last is with the sensor on the back of the wrist. Furthermore, we will assess whether factors such as movement, the number of desaturation events and skin type impacts the signal accuracy.

5.3 Data Analysis

5.3.1 Classification Performance Metrics

We use a binary classification system in order to classify desaturation events in the data. That is, we classify sections of the data as either a desaturation event or not a desaturation event. The data is divided into 60 second windows with SpO_2 signals and one label. Each predicted label will be compared with the actual label we get from the automatic scoring in Noxturnal. The window can then be classified as True Positive (TP) if both predicted and actual labels match on labelling a desaturation event, True Negative (TN) if they both match on there not being a desaturation event, False Positive (FP) if there is predicted a desaturation event but not an actual event, and False Negative (FN) for the reversed. This evaluation will be done on the data set from Nox and from Garmin, where the performance of both will be compared afterwards.

The performance of the binary classifiers will be evaluated by the four metrics accuracy, specificity, sensitivity, and Cohen's Kappa (κ). The first three metrics can be found in the confusion matrix in Figure 5.2.

Confusion Matrix

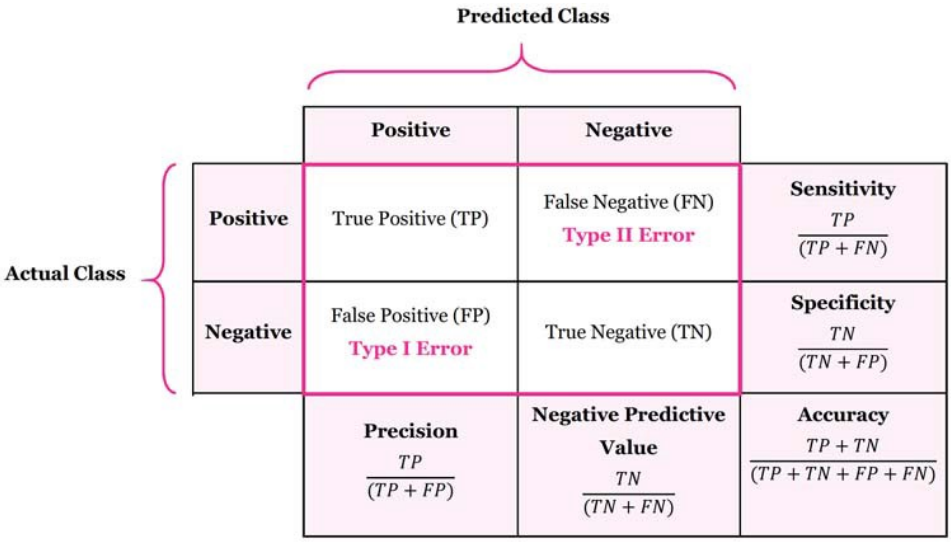


Figure 5.2: *Confusion matrix*

As we are using a binary classification system we can use a confusion matrix to measure the performance of the classifiers. A confusion matrix can be defined as a summary of prediction results on a classification problem [8]. We predict the outcome of a classification on a set of data, and compare this with the actual outcome. We summarize the predictions in the aforementioned different classes of TP, TN, FP, and FN by counting each prediction that falls under each class. From this, we can calculate several metrics that tell us something about our classifier.

From the confusion metrics we will use the metrics of accuracy, specificity and sensitivity. The equations for calculating these can be seen in Figure 5.2. To understand what the values we get are we will describe the equations with words. Accuracy takes all *correct* classifications and divides it on *all* classifications. Sensitivity takes all true positives, meaning actual desaturation events that were predicted, and divides on *all actual* positive values, meaning desaturation events that were predicted or not. Specificity does the same as sensitivity but for the negative values. That is, it takes all the true negatives and divides it on all actual negative values, those that were predicted and those that were not.

From the confusion matrix we also see there are two types of errors that can be made which are referred to as Type I and Type II errors. A Type I error is falsely classifying a negative as a positive, while Type II errors are falsely classifying a positive as a negative [8, 53]. How much we weight making each of these errors depends heavily on the domain. For instance, a model that continuously falsely predicts a person does not have cancer when in reality they do (Type II) can be fatal.

Cohen's Kappa (κ)

Cohen's Kappa (κ) [14] is a statistic that measures inter-rater and intra-rater reliability. It is a more robust metric than accuracy because it measures of how well the classifier performed compared to how well it would have performed by chance. κ is calculated as given in Equation 5.1 where P_o is the proportion of agreement and P_e is the proportion of agreement expected to be by chance. The upper limit of κ is 1, meaning it is a perfect agreement, while the lower can either be 0 meaning the agreement is by chance or between 0 and -1 for less than chance agreement. Cohen suggests that a value of κ larger than 0.0, 0.2, 0.4, 0.6, 0.8 and 0.9 should be interpreted to indicate none, minimal, weak, moderate, strong and almost perfect degrees of agreement, respectively [42].

$$\kappa = \frac{(P_o - P_e)}{(1 - P_e)} \tag{5.1}$$

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \tag{5.2}$$

In regards to binary classification and the confusion matrix, κ can be calculated as in Equation 5.2.

5.3.2 Signal Quality Metrics

Accuracy

Accuracy root mean square (A_{rms}) is the standard metric that is used to evaluate the quality of oximeter sensors [30]. It is expressed as the root

mean square between the pulse oximeter under test and the reference CO-oximeter. The formula for calculating A_{rms} is given in Equation 5.3, which is the same as root mean square error (RMSE). The value that we get from this describes all the points in the data set, as opposed to just one point. The closer the number we get is to 0 means the better accuracy as it indicates less error between the measures. In our case the Nox is the reference sensor while the Garmin watch is the test sensor.

Because outliers have an excessive negative effect on results of the A_{rms} parameter, we also assessed the mean absolute error (MAE), a measurement that is more robust in the presence of outliers [27]. This is because for MAE errors are weighted equally, while for RMSE they are wighted exponentially. That is, larger errors are weighted more than smaller errors. Willmott and Matsuura [77] also argued for the use of MAE in place of RMSE because of the ambiguity of RMSE. A later article by Chai and Draxsler argued against the complete avoidance of RMSE [11]. Equation 5.4 shows the formula for calculating MAE.

$$A_{rms} = \sqrt{\frac{\sum_{i=1}^{n} (SpO_{2i} - S_{ri})^2}{n}} \tag{5.3}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{5.4}$$

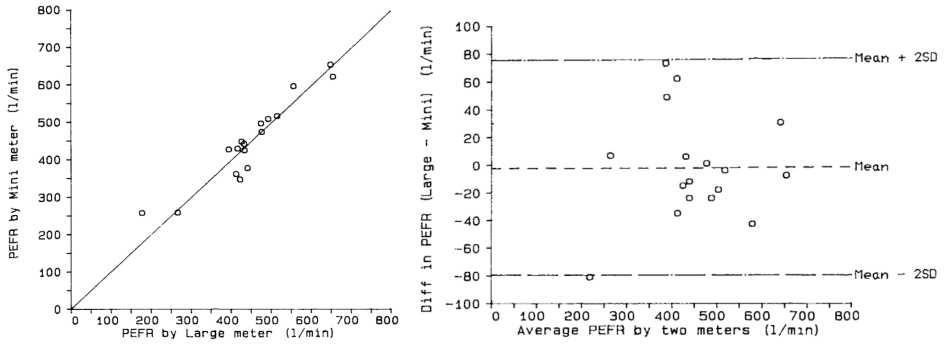
Bland-Altman analysis

In addition to accuracy, we will also perform what is known as a Bland-Altman analysis. In statistics when we work with two different variables we usually use correlation coefficient to assess their relationship. A correlation coefficient is a value between -1 and 1 which tells us the strength and direction of the relationship between variables [15]. A common measure is the product-moment correlation coefficient, also known as Pearson's r. While the correlation coefficient is used for assessing the relationship between two variables, it was found to not be ideal for measuring the agreement between two different measurement systems.

Martin Bland and Douglas Altman proposed what is known as the Bland-Altman analysis that addresses the issue of comparing one measurement to another [7]. Oftentimes when a new measurement method is evaluated, it is compared to the established or gold standard of measurement. If the agreement between the methods are sufficient, the new method can be used.

The Bland-Altman analysis is a graphical method that consists of a mean bias and upper and lower Limits of Agreement (LoA). A scatter plot is drawn where the x-axis is the mean between the two measurements ($(measurment_A + measurment_B)/2$) and the y-axis is the difference between the two measurements ($measurment_A - measurment_B$) [16]. The mean bias

is the mean of the difference(y-axis), while the upper and lower LoA are calculated by a 95% confidence interval of this bias. A complete agreement would give a mean bias of 0.



(a) *correlation example* **(b)** *bland-altman example*

Figure 5.3: *Examples of plots of the same data set [7]*

In an article by Doğan [16] some pitfalls in Bland-Altman analysis are addressed. For instance, there is an assumption that the differences need to be normally distributed. If this is not met, then the data may be logarithmically transformed. The differences should always be verified for normal distribution first, this by drawing a histogram or performing a Shapiro-Wilk test [22]. Another problem is that the sample size should be large enough to be universally valid.

5.3.3 Statistical Analysis

A lot of new data is generated from both the classification evaluation and quality evaluation of the raw data. These are in the form of metrics for each data set and also aggregate metrics for different subsets of the data. To make sense of the new data we use various statistical methods for description and summary. The following subsection is based on Britannica unless otherwise specified.

We use many different numerical measures to summarize for different purposes. The most common and widely used statistic is the mean or average for central tendency of a set of continuous values. We also use Standard Deviation (SD) for the spread around the mean and a 95% Confidence Interval (CI). To describe proportion of data in a category we use percentage. Lastly, we use quartiles to divide the data into percentiles of four equal parts from smallest to largest. Inter-quartile range is used to calculate lower and upper bounds in order to remove outliers of the data.

Hypothesis Testing

Hypothesis testing is defined by Britannica [65] as a form of statistical inference where we use data from a sample to draw conclusions about the population it is drawn from. Hypothesis testing starts with an assumption about the population, often called a null hypothesis (H_0). We also have an

alternative hypothesis (H_A) about the same population which is different, often opposite, from the null hypothesis. This is the hypothesis, or claim, we want to test. The procedure requires drawing a sample from the defined population and determining whether H_0 can be rejected or not. If it is rejected, then the conclusion drawn is that H_A is true.

There is a possibility that we falsely reject H_0. Correspondingly, there is a chance of falsely accepting H_0. These are called Type I and Type II errors respectively. We set a significance level, called α for the probability we would like to allow for making a Type I error. The common significance levels are $\alpha = 0.05$ and $\alpha = 0.01$. An observed level of significance called p-value is measured, and H_0 is rejected whenever it is smaller than the already chosen α.

Regression and Correlation

We conduct hypothesis testing in regression and correlation analysis to determine if a relationship is statistically significant. With regression we have a dependent variable and one ore more independent variables that we would like to determine the relationship between. In our case with signal quality, we assess if there is a relationship between the dependent variable of accuracy and some external factors. This is based on correlation.

A correlation coefficient gives us a value between -1 and 1 indicating how strong two variables are related. Either end means there is a strong positive or negative correlation respectively, while 0 means there is no correlation. The most common correlation coefficient is Pearson's product-moment correlation coefficient, otherwise known as Pearson's r. If the relationship is ordinal or not linear, an alternative coefficient is Spearman's *rho*. It is important to understand that correlation does not mean causation.

Statistical Tests and Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) test is a statistical procedure that is used to compare the means of two or more groups [4, 65]. The procedure uses a hypothesis testing in regards to the equality of the means between the groups. There are different version of the procedure related to if there is just one variable or multiple, and even if there are interactions between variables. It uses an F-test to check for equality for multiple groups, or a t-test for comparing two groups. ANOVA is a parametric test is based on some assumptions on the data. That is, that the residuals are linear, homoscedasticity or homogeneity of the variance between the groups and that the observations are independent.

We can use statistical test to assess these assumptions. For linearity this can be done with a Shapiro-Wilk test or a histogram, and homoscedasticity can be tested with a Brown-Forsythe test. If these assumptions are not met, we perform non-parametric tests like Wilcoxon rank sum test for comparing two groups or Kruskal-Wallis test for more than two groups instead of ANOVA.

5.3.4 Graphical Analysis

We use many different plots to visualize and make sense of the data. The plots used and their purpose are listed as follows:

Time series graph

A time series graph is a line graph where a value is plotted on a timeline. The graphs visualize the sensors oxygen saturation recordings over time. The level of oxygen saturation is plotted on the y-axis while the time is represented by the x-axis. A line is drawn between the plotted SpO_2 values signifying the continuous nature. The graphs for the reference and test oximeter are plotted in the same plot, allowing us to see where they match and deviate and by how much. It also allows us to visually adjust the graphs for better synchronization.

Histogram

The histogram gives a graphical representation of a data set by grouping values in equal sized bins of a specified range. This gives a visual representation of the frequency in each bin. As an example, we collect signal data with values that can range between zero and 100. Plotting the recorded signals in a histogram makes it easier to see the distribution of the signals.

Boxplot

Another way of visualising the distribution of data is through a boxplot, also called a box and whisker plot. A boxplot is a form of descriptive statistics that visualizes how the values in a data set are spread and skewed. It consists of a box which represent the 25th and 75th percentile of the data set, also known as first and third quartile. the 50th quartile is the middle value, also called median, in the data set and is represented by a line inside the box. From the box there are whiskers at the top and bottom which represent the 0th and 100th percentile of the values in the data excluding outliers. Outliers are plotted as dots outside of the plot.

Scatter plot

A relation plot, which is commonly referred to as a scatter plot, presents the relation between two separate sets of data. The data sets share a common relation, often time, and are plotted with one value on the x-axis an the other on the y-axis. In our case, the signals from the devices could be plotted against each other, or for accuracy could be plotted against the measured covariate e.g. movement. Scatter plots can also be used to visualize correlation by fitting a regression line to the data.

Bland-Altman plot

A Bland-Altman plot is a variation of a scatter plot. The difference is it plots the agreement between two methods rather than the correlation. How the same data set is visualized in these two plots can be seen in Figure 5.3. The plot graphically displays the mean bias of the method or measurement under test to the reference method. In our case the method under test is Garmin's pulse oximeter with Nox T3 being the reference oximeter. The mean bias is calculated by plotting the average between the devices to the difference. Both the upper and lower LoA are marked in the plot as well. These limits represent the 95% confidence interval of the bias.

QQ-plot

We test the data sets for normality by plotting the data in a QQ-plot. A QQ-plot, which stands for quantile-quantile plot, is a probability plot that compares the distribution of two data sets by plotting their quantiles against each other. If they are normally distributed the quantiles will lie in a 45 degree line.

Chapter 6

Implementation

In this chapter, we describe some of the code that is used. This includes for the app (Section 6.1) and the scripts that were created for preprocessing the data, classification, and signal quality calculations (Section 6.2). Some of the code is included in this chapter while the full source code is linked in Appendix A.

6.1 Real Time App

The full description of the app's implementation can be found in Halvorsen's thesis [26]. Some changes were made to the original implementation to accommodate the updates in the Companion SDK. These were however not extensive.

6.1.1 System Environment

The app is built in Android Studio with with the implementation written in Java and XML for design. The app was tested on an Android 10 phone, which was also the phone used during recording. The target SDK version was 31 while the minimum SDK was 21.

Software	**Version**
Android Studio	Bumblebee | 2021.1.1
Gradle	7.2
Java	1.8
Companion SDK	3.0.2
Android phone	10

Table 6.1: *Relevant software for the app*

6.1.2 Code Updates

Halvorsen's real time app was last updated early 2020. Since then, the Companion SDK has undergone some updates. Because of this the app

does not run in its current state. After inspection of the code and updates in the Companion SDK, the change that needed to be made was removing the `RealTimeDataManager` variable and adding the listener directly in the `DeviceManager`.

```
1  private boolean pulseOxListener(String device) {
2      DeviceManager devMgr = DeviceManager.getDeviceManager();
3      Device d = devMgr.getDevice(device);
4      listener.setStartTime(DateTime.now());
5
6      storage.putWriters();
7      storage.insertFirstRow();
8      listener.setStorage(storage);
9
10     d.samplingFrequency(TWENTY_FIVE_HERTZ);
11     //RealTimeDataManager rtMgr = devMgr.getRealTimeDataManager
           (); replaced with DeviceManager
12     EnumSet enumSet = EnumSet.noneOf(RealTimeDataType.class);
13     enumSet.add(SPO2);
14     enumSet.add(HEART_RATE);
15     enumSet.add(HEART_RATE_VARIABILITY);
16     enumSet.add(RESPIRATION);
17     enumSet.add(ACCELEROMETER);
18     devMgr.addRealTimeDataListener(listener, enumSet); //
           replaced with DeviceManager
19     devMgr.enableRealTimeData(device, enumSet); // replaced with
            DeviceManager
20     Log.i("Listener", "Enabled");
21     return true;
22 }
```

Listing 6.1: *Changes made in app*

The code in Listing 6.1 is located in the `CollectActivity` with the purpose of writing recorded data to file and enable listening to real time data from the sensors. Other than this, the User Interface (UI) was updated by enlarging the font size.

6.2 The scripts

The purpose of these scripts is to make analysing a large quantity of data more efficient. The section start off with an overview of the software used in the scripts before describing the different scripts.

6.2.1 System Environment

The scripts were implemented in *Python* as there are many available libraries that have the functions that are needed. We use the *Pandas*

library for all of the CSV-file and data handling, including resampling, interpolation, and dropping of columns. With *NumPy* we have access to arithmetic functions specialized for matrices. *Sci-kit Learn* and *SciPy* are scientific libraries extensively used for ML. All the ML classifiers are implemented with *Sci-kit Learn*. *Matplotlib*, *seaborn* and *pyCompare* are used for plotting the data into scatter plots, histograms, boxplot, Bland-Altman plot, to name a few. Statsmodels is used for statistical tests and models. Table 6.2 gives an overview of the most important software and libraries and the versions used for the scripts. All the required libraries can be found in a 'requirements.txt'-file in the source code.

Software	**Version**
Python	3.9
Anaconda	4.10.3
Pandas	1.4.1
NumPy	1.22.2
Sci-kit Learn	1.0.2
Matplotlib	3.5.1
Seaborn	0.11.2
SciPy	1.8.0
pyCompare	1.3.2
Statsmodels	0.13.2

Table 6.2: *Software used for the scripts*

6.2.2 Preprocessing

The data from both Venu 2S and Nox T3 are stored as CSV-files. A file called "preprocessing.py" is the first step for the raw data. Running the script requires five command-line arguments, such as the following example: `python preprocessing.py 1 0 1 -3 garmin`. The three first digits refer to the id, experiment type and iteration for the recording to be processed respectively. This is used to open the corresponding CSV-file and read it into a `pandas` dataframe. The fourth and last digit is the delay and the last argument is which device the shift of delay will be applied on. A sample of the "preprocessing.py"-code in which only the processing of Garmin signals is included is given in Listing 6.2.

```
1 # Read oximeter CSV-files and rename columns
2 garmin = pd.read_csv(f'./data/Sub00{id}/{ex}/Sub{it}{ex}{id}
      _Venu_2S_log_spo2.csv',
3                      sep=",",
4                      header=0,
5                      usecols=[1, 2],
```

```
6                   index_col=['Duration'],
7                   parse_dates=[0],
8                   date_parser=parseNow)
9 garmin = garmin.rename(columns={'Value': 'signal garmin'})
10
11 # Resample
12 garmin = garmin.resample("1000ms").mean()
13
14 # Interpolate
15 garmin = garmin.interpolate(method='quadratic')
16
17 # Remove outliers and NaN's
18 garmin = garmin.drop(garmin[garmin['signal garmin'] > 100].
    index)
19 garmin = garmin.dropna()
```

Listing 6.2: *Data processing of Garmin raw data*

After the data is downloaded into a `pandas` dataframe, we first rename the columns, resample the signals to 1Hz then we interpolate. Then outliers are removed, which in this case is SpO_2 values greater than 100, and also NaN values are removed. All these are from functions in the `pandas` library. After the SpO_2 signals have been synchronised they are added to a new dataframe where any additional signals have been trimmed away in order to equalize the lengths of the data sets. The preprocessed data is stored as a new CSV-file consisting of the columns 'signal garmin', 'signal nox', 'desat', 'movement' and 'artefact' in the 'processed' directory.

Synchronization

The SpO_2 signals are synchronized by finding the delay between the peaks in the SpO_2 signals. This means we assume that the peaks in the data reflects the same breath in real time. The script for finding the delay between two signals originally written by Løberg [40] can be seen in Listing 6.3. It utilizes the correlate function from SciPy that returns an array of correlation values for all possible alignments of the signals. The `findDelay(a, b)`-function returns the delay between the signals a and b as an integer.

```
1 def findDelay(a, b):
2     return (len(b) - 1) - np.argmax(signal.correlate(a, b))
```

Listing 6.3: *Finding delay between data by Halvorsen*

Sometimes there are no peaks in the SpO_2 data. On these occasions this form of synchronization is not optimal. For this reason we also use the `findDelay(a, b)` on the signals we have from both accelerometers. The last method is then to visually inspect the line graphs for the SpO_2 signals and shift according to what visually matches or gives better A_{rms}.

6.2.3 Event Classification

Before we can use the classifiers on the data sets, there is some more preprocessing needed. As we want the data set to be as close to representative of SA, we want there to be a minimum of two desaturation events in the data set. This also satisfies the requirement of a minimum of 200 signal samples. We also use a function for reshaping the signals into 60 second windows with one label. The label for the 60 second window is based on whether any of the original labels in the 60 second window was a desaturation event (label = 1). If it was, then the signals in that window are stored as a list with the integer 1 as label. Otherwise the label is 0.

ODI classifier

The script for detecting desaturation events was originally written by Halvorsen [26]. A modified version can be seen in Listing 6.4. This script is based on the ODI definition and is therefore referred to in this thesis as the ODI classifier.

```
1  def desaturation(data):
2      full = np.mean(data)
3      last = np.mean(data[len(data)-10:])
4
5      if ((full-last) >= 3):
6          return True
7      return False
8
9  def label_events(fullsignal):
10     total_desats = 0
11     ongoing = False
12     start = 0
13     events = []
14     labels = [0]*110
15
16     for i in range(len(fullsignal) - 120):
17         value = desaturation(fullsignal[i:i+120])
18         labels.append(float(value))
19         if not ongoing and value: # found start of desat event
20             start = i
21             ongoing = True
22             total_desats += 1
23         elif ongoing and not value: # found end of desat event
24             events.append((start+110, i-start))
25             ongoing = False
26
27     #print_events(events)
28     for _ in range(10):
29         labels.append(0)
```

```
30     return labels
```

Listing 6.4: *Modified algorithm for counting desaturation events based on the ODI definition*

The script has two functions; a short one `desaturation(data)` and a longer one `label_events(fullsignal)`. The first function takes in a 120 second window and looks for a desaturation in this window by comparing the mean value of the full window with the mean value of the last ten seconds. If the difference between these two is bigger than 3%, then the function returns *True*. This means a desaturation event was found, otherwise the function returns *False*. The longer function does four things, first it passes 120 second intervals to `desaturation(data)`, it keeps track of the number of desaturation events, the length of the desaturation events and where they occur, and lastly, it returns a list of the labels.

The longer function takes in the one parameter $fullsignal$. *Fullsignal* is the SpO_2 recording for a given data set. There is a for-loop that moves one second at a time across $fullsignal$. The `desaturation(data)` function is called within this loop and thus given a new window every second. The value that is returned by this function is stored in the variable *value* and also appended to the list *labels*. If `desaturation(data)` returns `True` then a desaturation event has been found. The event is ongoing (`True`) until the function returns `False`. When `desaturation(data)` returns `False`, that means it's the end of the event and it is now possible to find new events. The event meta data is stored as a tuple in the *events* list.

The original script has been slightly modified in this thesis. For instance, the longer function used to only find desaturation events. Now, it creates and returns a list of each labelled second.

We have created a script that loops through all the processed recordings, labels the data sets and then calculates the metrics of accuracy, sensitivity, specificity and κ. After the data sets, both Nox and Garmin, are labelled by the ODI classifier, they are reshaped into 60 second windows with one corresponding label. We only use data sets that have more than two windows of events. The prediction made by the classifier are then compared to that of Noxturnal by the four metrics. The results from the metrics are stored as a CSV-file in an 'output' directory.

ML classifiers

For the ML classifiers we need to balance the data before training and testing the data sets. This is because there is likely to be significantly more windows without events than with events. With an imbalanced data set we face the problem of the model always predicting the larger class, leading to inaccurately high accuracy.

```
1 def balance(signals, labels):
2     numApneic = len([l for l in labels if l != 0])
```

```
    numNormal = len([l for l in labels if l == 0])
    normalIndices = np.where(labels == 0)[0]
    apneicIndices = np.where(labels != 0)[0]
    if numApneic <= numNormal:
        indexes = np.random.permutation(len(normalIndices))
        normalIndicesShuffled = normalIndices[indexes]
        normalIndices = normalIndicesShuffled[:numApneic]
    else:
        indexes = np.random.permutation(len(apneicIndices))
        apneicIndicesShuffled = apneicIndices[indexes]
        apneicIndices = apneicIndicesShuffled[:numNormal]

    dataToInclude = np.sort(np.concatenate([normalIndices,
        apneicIndices]))
    signals = signals[dataToInclude]
    labels = labels[dataToInclude]

    numApneic = len([l for l in labels if l != 0])
    numNormal = len([l for l in labels if l == 0])

    return signals, labels, numApneic, numNormal, dataToInclude
```

Listing 6.5: *Balancing data sets*

The function `balance(signals, labels)` in Listing 6.5 takes in the signals and labels and counts the number of labels with desaturation and without desaturation. It balances the data for each class by random subsampling of the majority class. In more detail this means that the whole of the minority class will be included, while we choose at random the same number of events from the majority class. The function then returns the balanced signals and labels, the number of apneic and normal events, and a list of the indices included in the balanced data set.

The process for ML classifiers is similar to that of the ODI classifier although slightly different because we train a model so we can make predictions on a data set. We use three basic ML classifiers (KNN, RF, SVM) which are all implemented with the Sci-kit Learn library. The model is trained on all the data sets, excluding test set, after they have been reshaped into 60 second windows and balanced. When it comes to classifying the test set, we use the `fit` function the classifiers have. The predicted label is then compared to the actual label and we calculate the four metrics κ, accuracy, sensitivity and specificity. Lastly, we store the results in a CSV-file.

All three classifiers in wrapper functions can be seen in Listing 6.6. There is no hyperparameter tuning, instead the choice of hyperparameter for the models is based on previous results of ML classification on SA data [35, 36]. For the KNN classifier we use $k = 10$ for the hyperparameter and the weights parameter is set to distance. SVM has the sci-kit learn standard rbf as kernel and we use 200 trees in the RF classifier.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn import svm
from sklearn.ensemble import RandomForestClassifier

def createKNearestNeighbours(hyperparams):
    return KNeighborsClassifier(n_neighbors=hyperparams, weights
        ='distance')

def createSupportVectorMachine():
    return svm.SVC()

def createRandomForest(hyperparams):
    return RandomForestClassifier(n_estimators=hyperparams)
```

Listing 6.6: *ML classifiers*

Cross-Validation (CV)

We split the data into train and test sets with two different methods, which are with a holdout test set and with a K-fold CV. For the holdout test split we test each individual subject recording on the rest of the data set, depending on whether the test recording is a lab or overnight recording. The script for hold-out testing is located in the file called "holdout_testing.py".

For K-Fold CV we use a k = 10 in the `KFold` function, where one fold is the test set and the remaining nine are used for training. With CV the holdout method is repeated k times with a different subset of the ten as the test set. We test various subsets of the data with CV, like the different experiment sets, all the lab recordings, and also all the recorded data. The script for 10-fold CV testing is located in the file called "crossval_testing.py".

6.2.4 Sensor Quality Script

For easy calculation of the metrics, we have created a script that calculates all the metrics for each recording. The script iterates through all the recording files after they have been processed. First, the file is opened and read into a dataframe. Accuracy, MAE, precision, mean bias, upper and lower LoA are calculated for the data set. In addition to these metrics, we also add other columns. These are percentage of movement, percentage of desaturation, skin type, experiment type and total length of the data set. We add percentage of movement and desaturation for when we test if there is a correlation between these variables and accuracy. The skin and experiment type are for easy grouping of the recordings.

```
# iterate over files in processed directory
for file in os.scandir(directory):
    if file.is_file():
        # read processed file to df
        df = pd.read_csv(file.path, sep=',', header=0, index_col
            =[0])
        ex = int(file.name[3:4])
        sub = int(file.name[4:-4])
        length = df.shape[0]
        skin = ''
        desat = round((df['desat'].sum() / length) * 100, 3)
        movement = round((df['movement'].sum() / length) * 100,
            3)

        if file.name[:-4] not in remove:
            all_df = pd.concat([all_df, df], axis=0)
            if ex == 3:
                overnight_df = pd.concat([overnight_df, df], axis
                    =0)
            else:
                script_df = pd.concat([script_df, df], axis=0)

        if sub in skin_type['dark']:
            skin = 'dark'
        if sub in skin_type['medium']:
            skin = 'medium'
        if sub in skin_type['light']:
            skin = 'light'

        # create a list of file metrics
        accuracy = arms(df['signal garmin'], df['signal nox'])
        ma_error = mae(df['signal garmin'], df['signal nox'])
        precision, bias, upper_loa, lower_loa =
            bland_altman_analysis(
            df[['signal garmin', 'signal nox']])
        file_metrics = [file.name[:-4], accuracy, ma_error,
                        bias, precision, upper_loa, lower_loa,
                            desat, movement, skin, experiment_type
                            [ex], length]

        # append list to a list of metrics
        metrics.append(file_metrics)
```

Listing 6.7: *Excerpt of script for calculating metrics for all data sets*

The code for how the file-iterating is done can be found in Listing 6.7. There is a total of 12 columns where the first one, recording name or ID, is

the index column. When the metrics have been calculated for all the data sets, the list of all the metrics is first converted into a dataframe for it then to be stored as a CSV-file in a directory named 'output'.

Part III

Evaluation

Chapter 7

Results

In this chapter, we will present the results of the tests that were performed. We begin in Section 7.1 with an introduction to the subjects that participated and their demographics. More detailed information in regards to SpO_2 data collection is presented in Section 7.4. Before this, we present the results for the tests related to the app, i.e. usability testing in Section 7.2 and the apps ability to detect connection loss in Section 7.3. This is followed by an evaluation on the classification of desaturation events for Nox and Garmin signals in Section 7.5. Lastly we present the results for Garmin's pulse oximeter's signal quality in Section 7.6.

7.1 Subjects

A total of 15 subjects were recruited for this round of experiments. Of the 15 subjects, three of them performed usability testing of the app as well as sensor testing. The demographics of the subjects are six of them were female and the remaining nine were male. In regards to skin tone four were dark, four medium and five light. More information in regards to the experiments they performed and data collected can be found in Section 7.4.

7.2 Usability testing

With usability testing of the app we hope to uncover weaknesses that can improve the quality of the app. This section will present the results of the conducted tests. It will start off with a summary before going more in-depth about participants, how the test went, the two tasks and subsequent recommendations from the participants in findings. Lastly, some of the suggested improvements will be implemented in the app.

7.2.1 Summary

The goal of the usability test was to see how easy or difficult the app is to use for a first-time user of the intended audience. Additionally, we wanted to uncover any underlying bugs that were missed, or changes that could improve usability. A total of three participants took part in the test. The

sessions lasted between 10 to 15 minutes where the participants where first presented with the purpose and goal of the usability test. The sessions were audio recorded and the participants were informed about this beforehand along with their right to end the session whenever they wanted. They had to complete two tasks in addition to answering some introductory questions and questions around their experience and thoughts about the app. A more in depth guide can be found in Appendix C. All participants successfully completed both tasks and the overall impression of the app was that it was simple and easy to use. There were however some shortcomings and bugs that negatively impacted their performance.

7.2.2 Findings

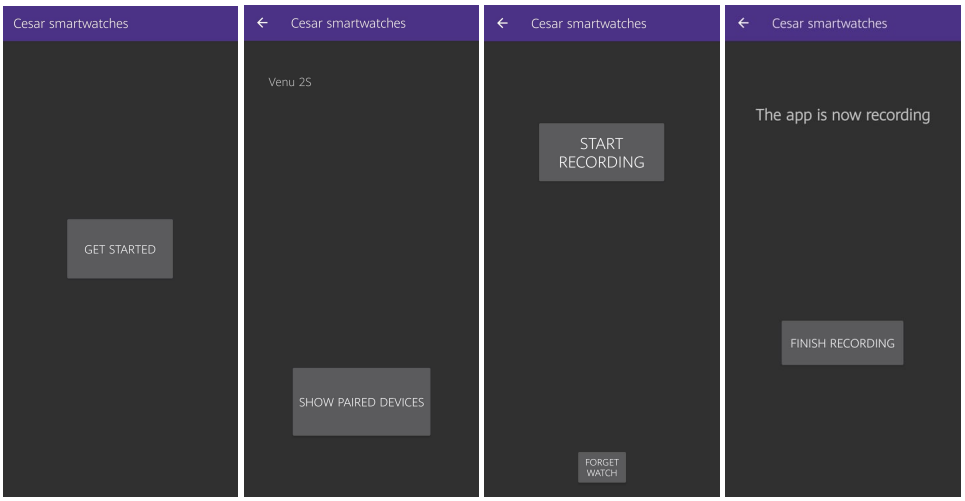
Participant	Age	Experience
1	20s	Never used a smartwatch before, but uses phone and apps regularly
2	20s	Currently has an Apple watch, previously had Garmin and Fitbit. Uses phone regularly and has apps connected to Apple watch
3	40s	Never used a smartwatch before, not that interested in apps. Uses phone mainly for calling

Table 7.1: *Summary of usability test participants*

The four screens in Figure 7.1 is how the app looked for the usability test. The size of the device name to be connected was enlarged since previous test. All the participants were first-time users of the app. A summary of the participants previous experience with smartwatches and apps can be seen in Table 7.1. Participant 1 is not familiar with smartwatches but is proficient with apps, Participant 2 is very familiar with smartwatches and apps connected to them while Participant 3 is not so proficient with apps or smartwatches. Participant 3 also has bad eyesight and doesn't always wear glasses while using the phone. All these participants successfully completed both tasks with varying degrees of support.

First task

For the first task the user had to pair the 'Cesar smartwatches' app with the Venu 2S smartwatch. All subjects made the same error on this task which was clicking on the 'Show paired devices' as can be seen on image b in Figure 7.1. They did not immediately understand that the device name was clickable and had to navigate back after voicing some confusion. For Participant 2 and 3 there was the additional bug that the device name did



(a) *main screen* **(b)** *pairing screen* **(c)** *start recording* **(d)** *recording*

Figure 7.1: *Four of the screens in the app before usability test*

not appear at first. The test manager had to exit and reenter the app for the name to appear.

The task is completed when the device name is visible on the 'Show paired devices' screen. It takes a few seconds from the device name is clicked until the pairing is completed. During this time the participants clicked the 'Show paired devices' expecting something to happen. This was a source of confusion for all participants as there was nothing visually confirming that pairing was happening or had finished other than the name almost randomly appearing on the paired screen. Because of this they all had to get verbal confirmation of the task being completed.

Second task

The second task was to start and end a recording on the app. The app is set back to the main screen and the participant would navigate from there. Since they had already developed some familiarity with the first part of the app, accomplishing this task was fairly easy for all the participants. The part that was different from the first task was choosing the paired device, press the 'Start recording' button and then the 'End recording' button to complete the task. There was no confusion around this part of the task.

Final impressions and recommendations

The general impression of the app from the subjects is that the app was easy to use. While Participant 1 found the simplicity a good thing, Participant 2 thought that it made the app boring and less professional. They all voiced the same sentiment that sometimes it was not clear what was supposed to be done, however the app was so simple that a few random guesses would complete the task. Things that were mentioned that could be improved are listed below:

Metric	Participant 1	Participant 2	Participant 3
Errors	1: Mistaken click of button	1: Mistaken click of button. 2: Crashing after device name click	1: Mistaken click of button. 2: Crashing after device name click
Satisfaction	Did not find it visually pleasing	Liked the simple design, straightforward and easy	Nothing special
Overall	Simple, easy to learn, confusing at times	Has the functionality that is required	Simple and confusing at times

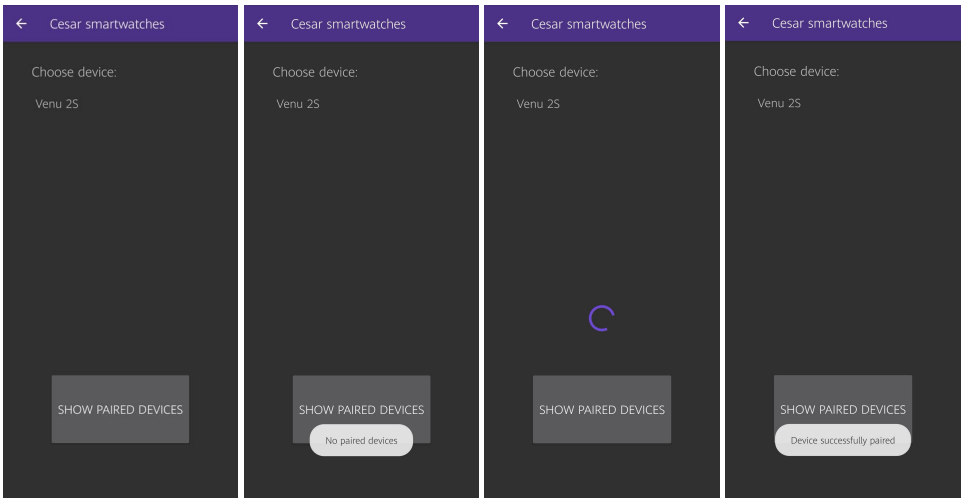
Table 7.2: *Summary of usability results*

- make it clearer that the listed device name is clickable,
- visualise the watch pairing/being paired,
- should not be possible to click the 'Show paired devices' button when there are no device paired,
- fix the app from crashing, and
- upgrade the design.

All in all the app should be more descriptive yet still keep the simplistic design.

7.2.3 Implemented Improvements

Based on the feedback from the test participants the most useful improvement was adding meaningful instructions to the app without causing too much noise. A text field has been placed over the devices list on both the pairing screen and paired devices screen instructing the user to pick a device. A progress bar was added to visualise the pairing. When the pairing is finished, a toast message will appear letting the user know if it was successful or not. Furthermore, clicking the "Show paired devices" button no longer leads to the next screen if no devices have been paired. The click will instead trigger a toast message informing the user to pair a device before they can move forward. The newly implemented changes can be seen in Figure 7.2.



(a) *pairing screen* **(b)** *no paired toast* **(c)** *progress bar* **(d)** *successful*

Figure 7.2: *Changes made to app after usability test*

7.3 Connection Loss Detection

The purpose of this experiment was to see if the results recorded by Halvorsen could be reproduced. A total of four experiments were done for connection loss detection, two for short duration, one for long duration and one testing close proximity with barrier. The length of short duration (1 min) and long duration (1 hr) is based on the times chosen by Halvorsen [26]. To ensure reproducibility and comparison with his results, we also use these times. For the experiment with a short duration we also tested the limits of the communication range. In order to measure the disconnection time we examine the 'Duration' column of the accelerometer data set as it has the highest sample rate of the sensors. If there has been a disconnect, there will be 'jump' in the data. By this we mean a gap in time between two timestamps.

7.3.1 Short Duration

For the short duration two different distances were tested. A shorter distance of 13 meters which is outside the communication range but not by far. This was included to see how sensitive the stated range is. Additionally, the watch and the phone were in two separate rooms with closed doors in between. For this shorter distance there was no disconnect.

The second distance was further, at approximately 20 meters, guaranteeing connection loss. With the longer distance there was a jump in duration. Accelerometer data was not registered between 26s and 109s during the recording as can be seen in Table 7.3. This is a total disconnect of 83 seconds which is longer than the intended 60 seconds. The results differ from the previously conducted test where the gap was shorter than the intended disconnect period.

TimeStamp	Duration
1644415994201	26
1644415994201	26
1644416077442	109
1644416077442	109

Table 7.3: *Connection loss for short duration*

7.3.2 Long Duration

For the long duration the disconnected period lasted for one hour. The experiment was conducted the same way where the recording app was left somewhere while the smartwatch was carried outside of communication range. The disconnection was registered between 129 and 3778 as seen in Table 7.4. This is approximately 60 minutes which means that there is no stored data between when the connection is lost and found again. The results also here differ from previous tests. The internal buffer or delay tolerance that was observed in the previous test is no longer present.

TimeStamp	Duration
1646751652210	129
1646751652210	129
1646755301822	3778
1646755301822	3778

Table 7.4: *Connection loss for long duration*

7.3.3 Close Proximity With Barrier

For this experiment the recording was started then the phone was placed 1-2 meters away. The person wearing the watch slept underneath covers with their body covering the wrist with the watch. A timer was set for five minutes. There was no disconnection for the accelerometer data meaning there was no disconnection between devices.

7.4 Data Collection

The watch could be worn three different ways for the lab tests; the comfortable way for the user, tight and with the watch face on the palm side of the wrist. Some performed all three while others only did one or two, and some performed the same twice. This resulted in a total of 34 recordings with the breathing script, 16 for normal wear, ten for tight wear

and eight for the last. The total duration of these experiments are 7 hrs, 23 min. We additionally had some overnight monitoring. A total of 12 recordings were performed lasting a total of 73 hrs, 43 min. The percentage of desaturation events ranged from 0-27.65% (mean: 8.67±7.7) for the lab test and from 0.1-14.36% (mean: 3.71±4.6) for overnight monitoring. A summary of which experiment was performed by each subject can be viewed in Table 7.5.

	Normal	**Tight**	**Back**	**Overnight**
Sub001	2	2	2	2
Sub002	1	1	1	2
Sub003	1	1	1	1
Sub004	1	1	1	0
Sub005	1*	1	0	1
Sub006	1	0	0	2
Sub007	1	1	1	0
Sub008	1	0	0	0
Sub009	1	1	0	1
Sub0010	1*	0	0	0
Sub0011	1	0	0	0
Sub0012	1	0	0	1
Sub0013	1	0	0	0
Sub0014	1	1	1	1*
Sub0015	1	1	1	1
Total	16	10	8	12

Table 7.5: *Subjects and performed experiments. *Not included in evaluation*

Each subject and recording combination is classified with the prefix 'R-' followed by some digits after they have been preprocessed and synchronized. The first digit denotes which recording it is as some subjects performed more than one recording. The second digit is for the type of experiment (normal watch wear = 0, tight wear = 1, back of the wrist = 2, overnight = 3) while the last digit(s) represents the id of the subject.

7.4.1 Errors

For some of the data collection not everything went according to plan. We present them here as they might have an effect on the results.

For some of the recordings the Nox oximeter and Venu watch were worn on different arms due to some confusion. This pertains to these recordings: first iterations for Sub001, Sub002 and Sub006, all recordings from Sub004, Sub005, Sub007, Sub008 and Sub009. The reason for this being an issue is that for synchronization we need to establish that the

sensors pick up the same data. Upon further examination of the oximeter data we see no major difference compared to the experiments where they were worn on the same arm. No data sets were therefore excluded based on this point.

Some subjects did not wear the nasal cannula. Since the signal we are most interested in is oxygen saturation which is recorded by the Nonin wrist oximeter, we assumed this would not be a problem. Additionally, we do not know how Noxturnal identifies desaturation events and which data it uses. Looking at the data afterwards however, we see that no events were registered for these. The affected data sets include R-1010, R-1011 and R-1012. There might have been other reasons for there not being any desaturation events such as individual differences, as there were other subjects who did wear the cannula without there being any desaturation events.

As the overnight monitoring was unattended, we relied on the subjects self-reporting of how the recording went. For the most part all the sessions went well, albeit some of them reported not sleeping as well as they normally would. On other occasions some reported more extensive issues. Subjects 1, 5 and 9 reported taking some of the equipment off during the night. For Subject 1 the finger probe was taken off for about five minutes, Subject 5 switched the finger the probe was on because of discomfort and Subject 9 had to rearrange the nasal cannula. The most extensive issue was Subject 14 not recording the signals from the smartwatch. There was some confusion around which app to use on the phone. We therefore have no corresponding recording for the watch from this session, only from Nox.

7.5 Desaturation Event Classification

This section presents the classification performance of four different classifiers on SpO_2 signals from Garmin and Nox. We start by first describing the data that was used for classification and how they varied for the four different experiment types. Then, we present the classifiers performance for holdout testing and 10-fold CV based on the metrics accuracy, specificity, sensitivity and κ. We visualize the performance with boxplots and compare the spread in the data with the mean. After this, we compare the performance of the classifiers and how it varied for the two devices. We then use statistical tests to determine if there are any differences between classifiers or devices. Lastly, we summarize the results and discuss how they contribute to answering our problem statement.

7.5.1 Classification Data

The goal of these experiments is to compare the performance of signal counting, referred to as the ODI classifier, and three ML classifiers (KNN, SVM, RF). We compare how they perform on signals from Nox and Garmin, and use the Noxturnal software's automatic scoring as the actual labels.

We use all recorded data that had two or more windows of desaturation

	Data sets	Apneic	Total	Balanced
Normal	8	41 (36.3%)	113	74
Tight	8	30 (27.0%)	111	60
Back	5	23 (34.3%)	67	46
Overnight	11	338 (7.7%)	4417	676
Total	32	432 (9.2%)	4708	856

Table 7.6: *The data sets and the number of apneic events used in classification*

events for training and testing the model. The training set consists of all these data sets while the test set is each individual data set. For each experiment the number of data sets that met this requirement were eight for normal, eight for tight, five for back and eleven for overnight, which is not much data. The number of windows and events for each experiment type can be seen in Table 7.6. The data sets are relatively imbalanced, which is the result of data being collected from subjects without SA and the subjects not being able to simulate enough events. For the ML classifiers the events are balanced via random sub-sampling of the majority class before classification.

7.5.2 Holdout Test

Table 7.7 presents the means of all the metrics for each classifier and experiment combination for both the devices. The performance for each individual data set can be found in Appendix D. For the ODI classifier there is no prediction being made, so the values calculated are definite. In contrast, the ML classifiers use each data set as a holdout test set while the rest of the data (lab recordings for lab test set, overnight recordings for overnight test set) for training the model. The predictions made are therefore based on the data the model is trained on.

The initial observation from Table 7.7 is that the classifiers perform on average better on Nox data compared to Garmin data on all metrics. This would be expected based on the fact of better data quality from a medical grade sensor where noise artifacts are removed, and potential preprocessing in Noxturnal. The same was not the case for the tight and back experiment for KNN and tight experiment for ODI, where the classifiers perform better on Garmin data based on κ and accuracy.

From the means we see that RF is the best performing classifier overall for Nox data with mean κ at 42.0%, while KNN is the best in regards to Garmin at 10.9%. A κ between 0.4 and 0.6 means the agreement is weak, while below 0.2 indicates no agreement. KNN has minimal agreement on Nox data with mean κ being 30%. The mean across experiments with the ODI classifier for Nox signals is 7.3% which means there is no agreement. SVM ranges from 17% for tight (none) to 47% κ overnight (weak). As KNN has the best κ on Garmin data of all the classifiers, we see there is no agreement between prediction and actual labels.

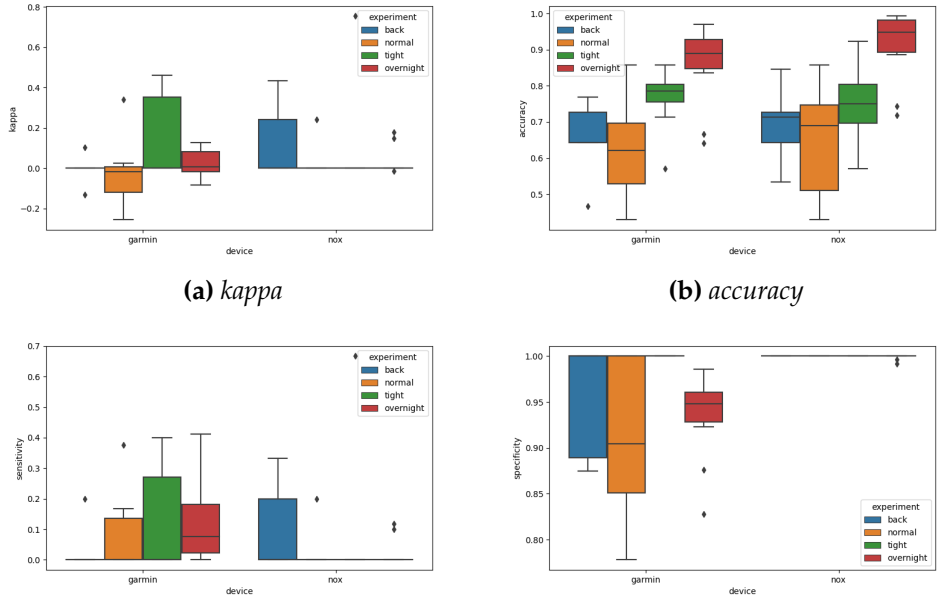
Holdout test set									
		κ		Accuracy		Sensitivity		Specificity	
Classifier	**Data**	**N**	**G**	**N**	**G**	**N**	**G**	**N**	**G**
ODI	Normal	0.03	-0.021	0.64	0.63	0.025	0.083	1.0	0.91
	Tight	0.094	0.15	0.75	0.77	0.083	0.12	1.0	1.0
	Back	0.14	-0.0058	0.69	0.65	0.11	0.04	1.0	0.95
	Night	0.028	0.028	0.91	0.86	0.02	0.12	1.0	0.94
KNN	Normal	0.32	-0.033	0.66	0.48	0.38	0.49	0.94	0.48
	Tight	0.17	0.19	0.59	0.6	0.31	0.64	0.86	0.55
	Back	0.24	0.32	0.62	0.66	0.28	0.5	0.96	0.81
	Night	0.47	-0.04	0.73	0.48	0.55	0.5	0.92	0.46
SVM	Normal	0.45	-0.017	0.72	0.49	0.64	0.53	0.8	0.45
	Tight	0.27	0.12	0.64	0.56	0.68	0.63	0.59	0.49
	Back	0.18	-0.034	0.59	0.48	0.51	0.47	0.67	0.5
	Night	0.54	0.05	0.77	0.53	0.76	0.39	0.78	0.66
RF	Normal	0.25	-0.038	0.63	0.48	0.56	0.42	0.69	0.54
	Tight	0.46	0.19	0.73	0.59	0.71	0.66	0.75	0.53
	Back	0.46	-0.055	0.73	0.47	0.54	0.4	0.92	0.54
	Night	0.51	0.11	0.76	0.55	0.72	0.58	0.79	0.53

Table 7.7: *Mean of metrics for all classifiers for each experiment and device combination. N = Nox, G = Garmin*

ODI has the best mean accuracy for both Nox and Garmin data (74.8% for Nox, 72.8% for Garmin), followed by RF on Nox data and KNN on Garmin data. The overall worst accuracy is 47% by RF, back experiment on Garmin data, while the best is 91% by ODI, overnight on Nox data. The sensitivity is below 0.1 for both devices for ODI while the specificity is above 0.9. Sensitivity tells us about how well the classifier identifies TP, in our case that is how well it detects desaturation events. On the other hand, specificity tells us how well it identifies TN, which is normal saturation. With sensitivity this low and specificity that high it means ODI accurately classifies TN while not detecting any actual events. All classifiers display the same trend with better specificity than sensitivity, though the discrepancy is not as big as with ODI. This trend might be due to the lack of desaturation events in the data, or the desaturation percentages not being that low compared to baseline.

In the following we will describe the performance of each classifier individually. The four metrics are plotted in boxplots to visualize the distribution of the metrics in each experiment. The boxplots are given in Figure 7.3 for ODI classifier, Figure 7.4 for KNN, Figure 7.5 for SVM, and Figure 7.6 for RF. For each plot we see the metric on the y-axis and device (Garmin or Nox) on the x-axis. There is a separate boxplot for each experiment type for both devices. It should be noted that the sample sizes

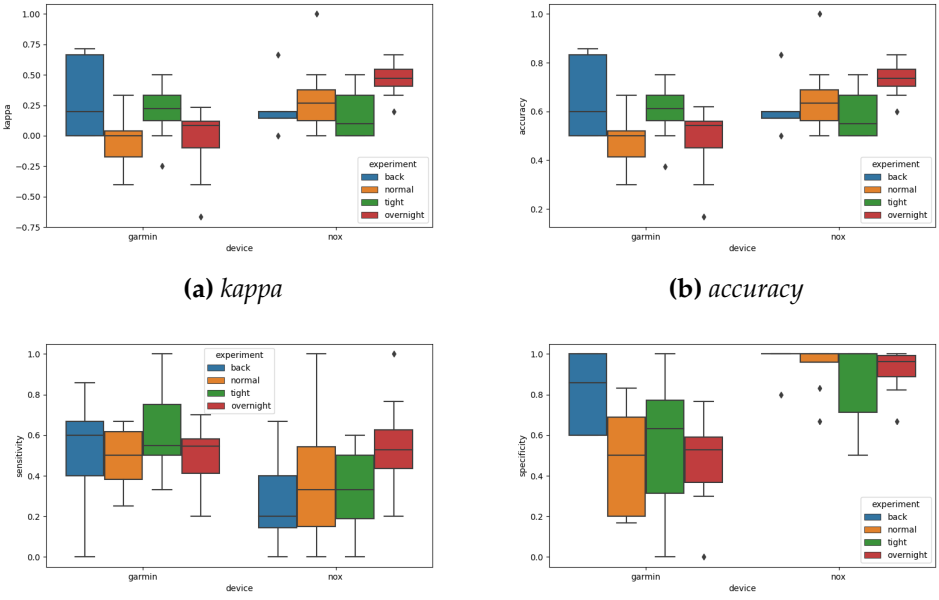
were not that large for each experiment type.



(a) *kappa* **(b)** *accuracy*

(c) *sensitivity* **(d)** *specificity*

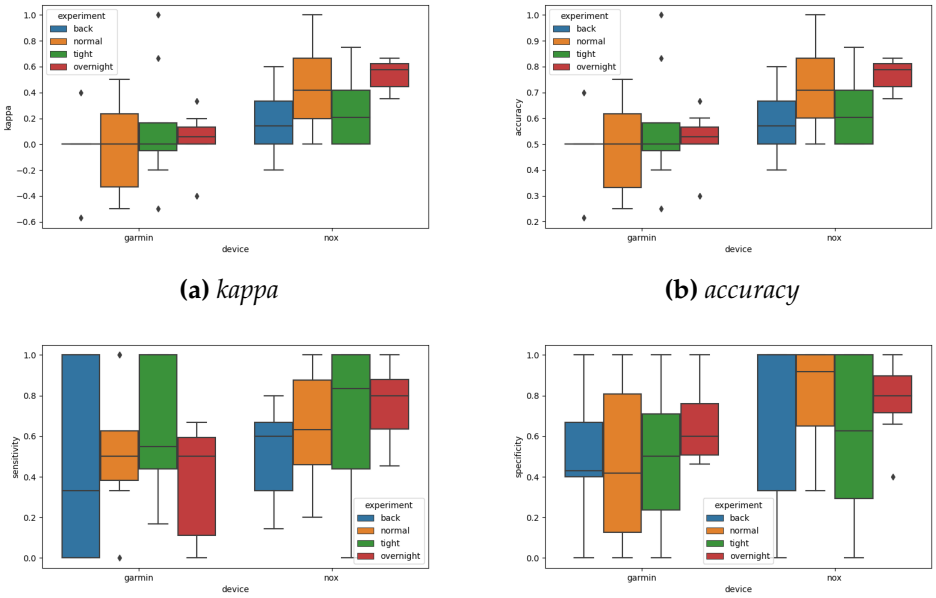
Figure 7.3: *Boxplots of metrics for ODI classifier*



(a) *kappa* **(b)** *accuracy*

(c) *sensitivity* **(d)** *specificity*

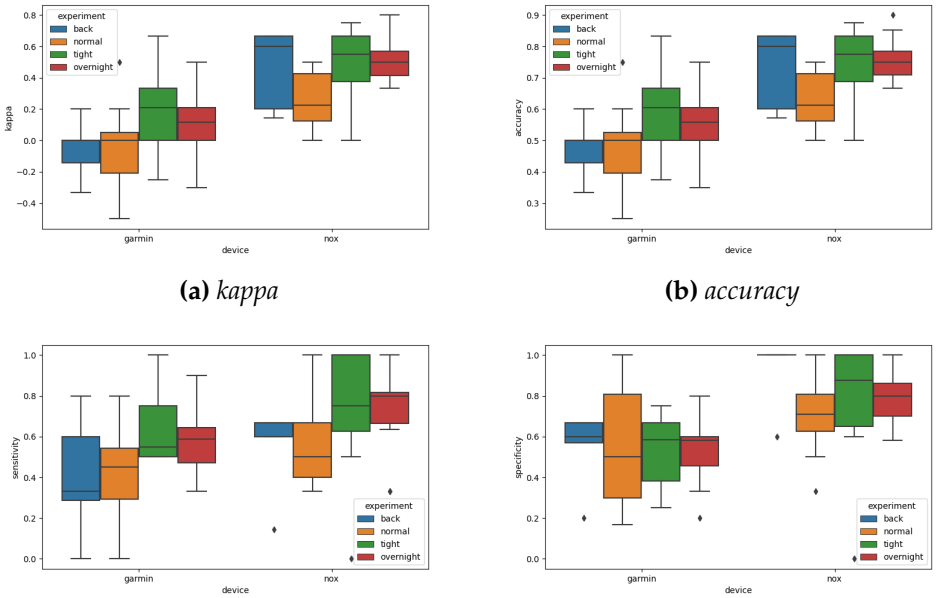
Figure 7.4: *Boxplots of metrics for KNN classifier*



(a) *kappa* **(b)** *accuracy*

(c) *sensitivity* **(d)** *specificity*

Figure 7.5: *Boxplots of metrics for SVM classifier*



(a) *kappa* **(b)** *accuracy*

(c) *sensitivity* **(d)** *specificity*

Figure 7.6: *Boxplots of metrics for RF classifier*

ODI

The accuracy, specificity, sensitivity and κ metrics visualised as boxplots for the ODI classifier is given in Figure 7.3. For each plot we see the metric on the y-axis and device (Garmin or Nox) on the x-axis. The κ for both devices

has the median equal to zero for all experiment variants. There is some spread for normal, tight and overnight variant for Garmin, and the back variant for Nox. Only the normal and overnight experiments for Garmin spread into negative values. The accuracy is best for overnight monitoring for both devices, while the worst is for normal experiment which also had the largest spread (0.4 - 0.8). The sensitivity is low for all experiments for both devices, with the experiments that had boxes and whiskers being positively skewed (right-skew). On the other end, specificity is high for all experiments for both devices.

Both devices were similar on the median values for all the metrics though the results were not particularly good. The median κ was at approximately 0.0 for all experiment types. The biggest spread was tight experiment however it was only to 0.4. Comparing the boxplots with the means given is the same.

ML Classifiers

A first observation for all the ML classifiers is that κ and accuracy have the exact same spread for the corresponding experiment and devices. The median is slightly better for Nox data than Garmin on these two metrics for SVM and RF.

A closer look at the boxplots for KNN shows a smaller range for κ and accuracy on Nox data, with overnight having the smallest range and therefore best performance. The largest spread for κ and accuracy was on the normal experiment ranging from approximately 0.0 to 0.75 for κ and from approximately 0.5 to 0.9 for accuracy. For sensitivity, both Garmin and Nox has large ranges in the boxplots, the worst being normal for Nox ranging from 0.0 to 1.0 with a slight positive skew (right skew). Specificity for Nox has a smaller range for back, normal and tight (0.8 - 1.0), though there is some outliers. On Garmin data the range is larger, here also a data set, tight, ranging from 0.0 to 1.0.

For SVM, the best performing experiment overall for both devices and on accuracy and κ is overnight. The highest κ scored is approximately 0.6 compared to 1.0 for normal. However, the range between smallest to highest value is better for overnight than normal, which makes it more consistent. Specificity for SVM ranges from 0 to 1 for all experiments with the exception of overnight for both devices and normal for Nox. This is also the case for back experiment on Garmin and tight experiment for Nox in regards to sensitivity. Such a big spread could mean that the classifier performs poorly because of the quality of the data.

RF has the best mean based on κ for Nox data, though the smallest value is low as 0.0 (normal and tight). The range is also smaller for Nox data than Garmin on all metrics. The largest range at 0.8 from smallest to largest value is for sensitivity on Garmin data (back and normal), and specificity on Garmin data (normal).

7.5.3 10-Fold Cross-Validation

In addition to the holdout test, we perform a 10-fold CV for subsets of the data. This is a preferred method in ML as it gives the model opportunity to train and test on different splits of the data. With this method we do not test the individual recordings, which means there might be data from different recordings in the test set. The disadvantage of this is that it no longer reflects the real world setting of individual subject diagnosis. We still include this as it makes it comparable with related work. We perform CV on the whole data set in addition to the subsets of the different experiments performed (normal, tight, back, overnight), and all the lab recordings. The results of 10-fold CV for Garmin data are given in Table 7.8, while the Nox results are in Table 7.9.

10-fold CV - Garmin					
Classifier	**Data**	κ	**Accuracy**	**Sensitivity**	**Specificity**
KNN	Normal	-0.16(±0.2)	0.45(±0.1)	0.36(±0.3)	0.5(±0.3)
	Tight	-0.03(±0.4)	0.48(±0.2)	0.4(±0.3)	0.55(±0.3)
	Back	0.02(±0.4)	0.5(±0.2)	0.68(±0.4)	0.37(±0.3)
	Night	0.03(±0.1)	0.51(±0.1)	0.42(±0.1)	0.61(±0.1)
	Lab	0.12(±0.2)	0.56(±0.1)	0.57(±0.2)	0.55(±0.2)
	All	0.06(±0.1)	0.53(±0.1)	0.5(±0.1)	0.56(±0.1)
SVM	Normal	-0.24(±0.3)	0.38(±0.1)	0.33(±0.4)	0.43(±0.4)
	Tight	0.02(±0.3)	0.53(±0.1)	0.58(±0.4)	0.43(±0.3)
	Back	-0.12(±0.4)	0.36(±0.2)	0.45(±0.4)	0.42(±0.4)
	Night	0.0(±0.2)	0.49(±0.1)	0.42(±0.2)	0.58(±0.2)
	Lab	0.13(±0.3)	0.56(±0.1)	0.61(±0.2)	0.52(±0.2)
	All	0.1(±0.1)	0.55(±0.1)	0.52(±0.1)	0.58(±0.1)
RF	Normal	-0.24(±0.2)	0.41(±0.1)	0.44(±0.3)	0.3(±0.3)
	Tight	-0.15(±0.4)	0.43(±0.2)	0.28(±0.3)	0.58(±0.3)
	Back	0.02(±0.3)	0.42(±0.3)	NaN	NaN
	Night	0.1(±0.1)	0.54(±0.0)	0.55(±0.1)	0.55(±0.1)
	Lab	0.04(±0.1)	0.52(±0.1)	0.5(±0.2)	0.54(±0.2)
	All	0.15(±0.1)	0.58(±0.1)	0.54(±0.1)	0.61(±0.1)

Table 7.8: *Mean and SD of metrics for all classifiers for different subsets of data*

On Garmin data, none of the classifiers got a κ greater than 0.2, the best being RF for the whole data set at 0.15(±0.1). The accuracy ranges from 0.36(±0.2) (SVM back) at the worst, to 0.58(±0.1) (RF all) at best. The results of the CV is similar to the holdout for Garmin data.

10-fold CV - Nox					
Classifier	**Data**	κ	**Accuracy**	**Sensitivity**	**Specificity**
KNN	Normal	0.21(±0.2)	0.59(±0.2)	0.29(±0.3)	0.92(±0.2)
	Tight	0.0(±0.3)	0.47(±0.2)	0.17(±0.2)	0.81(±0.3)
	Back	NaN	0.57(±0.2)	NaN	0.83(±0.3)
	Night	0.49(±0.1)	0.74(±0.1)	0.56(±0.1)	0.94(±0.0)
	Lab	0.23(±0.2)	0.62(±0.1)	0.31(±0.2)	0.92(±0.1)
	All	0.42(±0.1)	0.71(±0.0)	0.5(±0.1)	0.92(±0.1)
SVM	Normal	0.2(±0.3)	0.55(±0.2)	0.72(±0.3)	0.42(±0.4)
	Tight	0.21(±0.3)	0.57(±0.2)	0.46(±0.4)	0.76(±0.4)
	Back	0.27(±0.4)	0.6(±0.2)	0.62(±0.4)	0.68(±0.4)
	Night	0.55(±0.1)	0.78(±0.1)	0.68(±0.1)	0.88(±0.1)
	Lab	0.33(±0.2)	0.67(±0.1)	0.6(±0.2)	0.74(±0.2)
	All	0.51(±0.1)	0.76(±0.0)	0.63(±0.1)	0.88(±0.1)
RF	Normal	-0.08(±0.2)	0.49(±0.1)	0.37(±0.2)	0.53(±0.3)
	Tight	-0.17(±0.4)	0.42(±0.2)	0.35(±0.3)	0.46(±0.4)
	Back	0.15(±0.3)	0.59(±0.1)	0.49(±0.3)	0.69(±0.4)
	Night	0.57(±0.1)	0.79(±0.1)	0.77(±0.1)	0.82(±0.1)
	Lab	0.35(±0.2)	0.68(±0.1)	0.58(±0.2)	0.78(±0.2)
	All	0.52(±0.1)	0.76(±0.1)	0.74(±0.1)	0.78(±0.1)

Table 7.9: *Mean and SD of metrics for all classifiers for different subsets of data*

The classifiers performance is better on Nox data, the same as with the holdout tests. Additionally, we see that κ and accuracy is better for the larger data sets (overnight, all of lab, all the data) with the overnight data being the best overall. The best overall results for all metrics was by RF on overnight data. At the same time, RF also had one of the worst results which was for tight (κ = -0.17(±0.4)).

7.5.4 Comparing Classifiers

The best performing classifier is RF based on the mean of the metrics for holdout tests. More specifically, RF performed the best on Garmin data while KNN performed the best on Nox data. We will now compare the performance between the classifiers to see if there is any significant difference in performance. Upon further look at the distribution within each experiment type we see that many of the metrics have a large spread. Some of them range the full spectrum between 0 and 1 (-1 and 1 for κ). This is most likely due to the small sample sizes.

As the ODI classifier is an outdated classification method and did not even get a κ larger than 0.2 on Nox data, we will not include this classifier in the comparisons. Furthermore, because of the small size of samples for each experiment type, we will look at each classifier with the whole sample.

Based on the boxplots there seem to not be a significant difference between the experiment types.

We first perform Shapiro-Wilk tests to assess normality of the metrics for the holdout tests. Based on the p-values being significant ($p < 0.05$) for some of the metrics, we conclude that they are not normally distributed. For simplicity we use Wilcoxon rank sum test instead of t-test for all metrics. None of the classifiers differed significantly ($p > 0.05$) for the three metrics κ, accuracy or specificity. SVM and RF is significantly better than KNN at p-value = 0.013 and p-value = 0.0029 respectively.

7.5.5 Comparing Devices

From the boxplots and means of the metrics we see that the classifiers perform better on Nox data compared to Garmin. We investigate this further by performing Wilcoxon rank sum test between the devices for each classifier, also excluding ODI. κ, accuracy and specificity significantly performed better on data from Nox than Garmin ($p < 0.05$) for the KNN classifier while sensitivity got $p = 0.44$. The same was also the case for RF though the sensitivity got $p = 0.13$. For the SVM classifier κ and accuracy were significant ($p < 0.05$) while sensitivity got $p = 0.051$ and specificity got $p = 0.29$. In regards to κ and accuracy for all three classifiers being significant we can conclude that the performance on Nox data and Garmin data are different.

7.5.6 Comparing With Related Work

There are two previous works which are directly comparable to this by Kristiansen et al. [35, 36]. Both study the use of ML classifiers for apnea detection based on various signals, oxygen saturation being of most relevance here. Similarly to this, they also use 60-second periods of signals for labelling. It should be noted that the databases used in these studies are far larger than ours at 856 balanced periods and 4708 unbalanced.

In [35], artificial NN, SVM, decision tree, KNN, and RF were used for classification on data from two separate databases (Apnea-ECG with 3947 minutes, MIT-BIH with 76 to 3307 minutes). The accuracy was more than 90% for all classifiers with different signal combinations for the Apnea-ECG data, while the accuracy for the MIT-BIH data was in the range of 60-70% due to poorer data quality in the form of noisy signals. The latter is similar to our results on Nox data. In [36], 27 different classifiers were used on data from the A3 study (228,018 balanced periods). SVM was the worst classifier on oxygen saturation signals with κ of 0.31 (accuracy = 0.66), while the best classifier was GRUS with κ of 0.71 (accuracy = 0.85).

7.5.7 Summary

The goal of the classification evaluation was to see if ML classifiers are better at labelling desaturation events than signal counting. We additionally want to see how good the classifiers are at classifying events

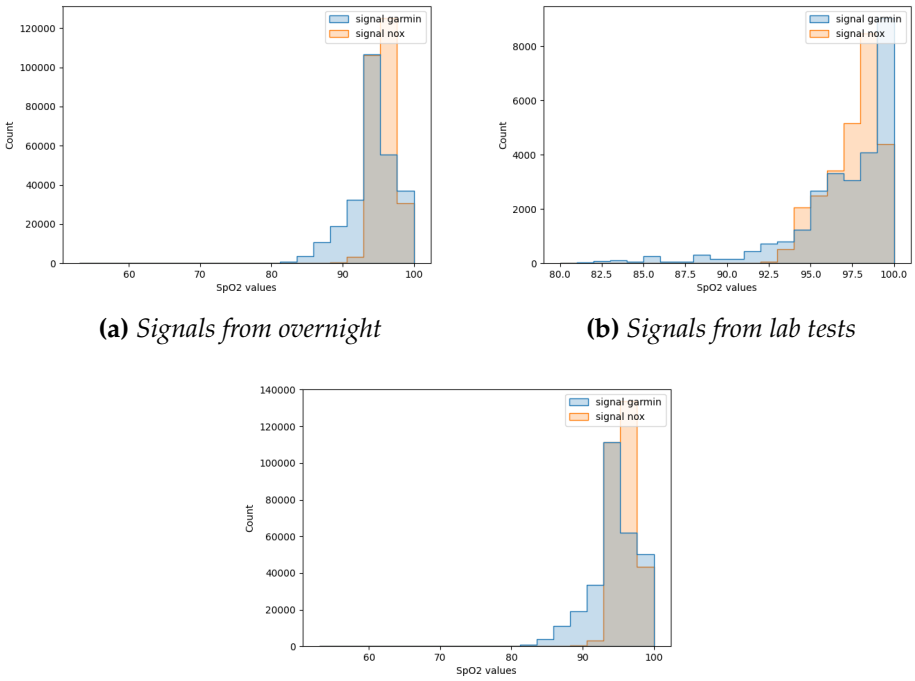
in SpO_2 signals. Furthermore, we wanted to investigate if SpO_2 signals from Garmin are as good as Nox in regards to event classification. We have evaluated four different classifiers based on four metrics. The data was tested by using both hold-out testing and CV. The results can be summarized as this:

- RF is the best performing classifier.
- The worst performing classifier is ODI, even on Nox data. The performance was poor with κ around 0.0 and is not comparable to ML classifiers. An explanation for this could be that it is strictly based on the definition, and it does not take into consideration that there might be noise or normal variations in the data set.
- All the ML classifiers performed better on the Nox data than Garmin data. This was significant based on a Wilcoxon rank sum test. The reason for this could be due to the quality of the collected signals. In the next section we will assess the quality of the Garmin pulse oximeter signals.
- The overall performance of the ML classifiers on the Nox data was similar to that seen in previous research. The same can not be said of Garmin, with accuracy ranging from 0.1 to 1.0 and κ in minus
- Due to the small sample sizes it is difficult to draw any definite conclusions from the results. This was especially evident from the results of the CV as the subsets with the experiments performed worse than the larger subsets or the whole data set.
- The most useful data collected was from overnight monitoring, unfortunately we only had 11 samples.

7.6 Signal Quality Evaluation

In this section, the quality of the sensor will be evaluated. The metrics they will be assessed on include A_{rms} for accuracy, MAE and we also perform a Bland-Altman analysis consisting of mean bias, precision (two standard deviations (SD) of the difference between devices), upper and lower LoA. The metrics are calculated on the preprocessed data. Furthermore, we have hypothesized whether some factors could explain the variation in sensor quality. The factors tested are the way the watch was worn, skin tone of the subject, number of desaturation events and movement in the data set. The section will start with the overall results of the quality assessment before the hypothesis testing.

7.6.1 Quality Results



(a) *Signals from overnight*

(b) *Signals from lab tests*

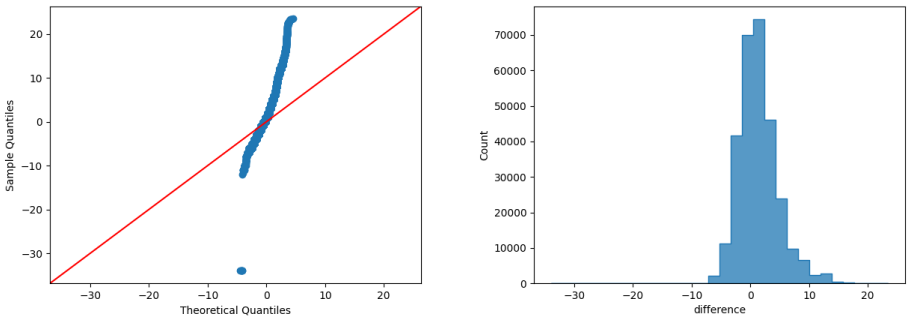
(c) *All signals*

Figure 7.7: *Histograms of the signals*

Figure 7.7c shows the distribution of signal samples from all the experiments for Garmin and Nox. The blue is the signals from Garmin and the orange from Nox, with the gray color representing the overlap of the two. Most of the samples came from overnight monitoring. From a total of 291 992 SpO_2 signals sampled from each device, approximately 99.9% of the SpO_2 values from Garmin are above 80%. For Nox 99.8% of the SpO_2 values are above 90%. Figure 7.9a more clearly displays the distribution of the signal values in a scatter plot. With regards to A_{rms} the accuracy

should be $\leq$ 3% according to the ISO standard for SaO_2 in the range of 70%–100%. The FDA, however, has specified typical accuracy values for reflectance oximeters as $\leq$ 3.5% which we will set as the upper threshold.

As we use Bland-Altman analysis for assessing agreement between the signals we first check if the assumption of normality in the difference between the signals are met. All signals are plotted in a histogram and a QQ-plot in Figure 7.8. For the data to be normally distributed the quantiles (blue dots) would lie in a 45 degree angle (red line). From Figure 7.8a we see this is not the case for our signals. The histogram of the differences also show that the data is not normally distributed as it is pointier than the signature bell-curve of a normal distribution. Despite the assumption of normality not being met we still use the Bland-Altman analysis and plot to assess signal agreement. The reason for this is that it mostly affects the LoA and not the mean bias.



(a) *QQ-plot* **(b)** *histogram*

Figure 7.8: *Plots for assessing normality in signal differences*

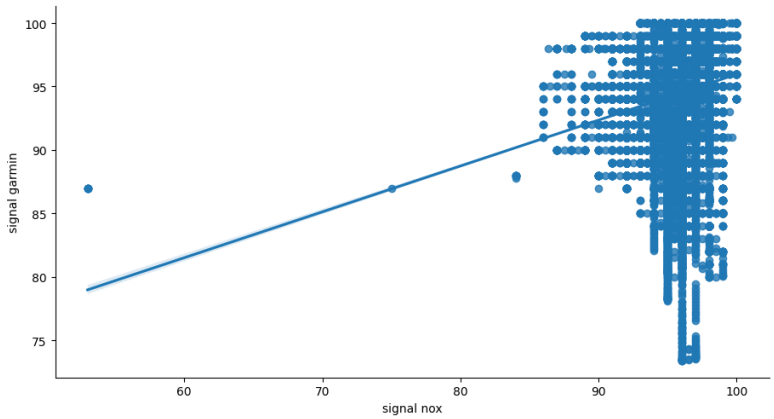
Summary Metrics						
	Accuracy	**MAE**	**Mean Bias**	**Precision**	**Upper LoA**	**Lower LoA**
All	3.718	2.718	1.391	6.896	8.149	-5.366
Mean	3.01(±1.6)	2.39(±1.3)	0.62(±2.1)	4.73(±2.3)	5.26(±4.1)	-4.02(±1.6)
Lab	3.363	2.308	0.361	6.688	6.915	-6.193
Mean	2.84(±1.8)	2.29(±1.4)	0.33(±2.2)	4.39(±2.5)	4.62(±4.3)	-3.97(±1.7)
Overnight	3.752	2.759	1.495	6.882	8.239	-5.25
Mean	3.48(±1.2)	2.69(±1.0)	1.48(±1.7)	5.72(±1.4)	7.08(±2.8)	-4.13(±1.2)

Table 7.10: *Mean and SD of signal quality metrics for all recordings and the subsets of lab and overnight recordings*

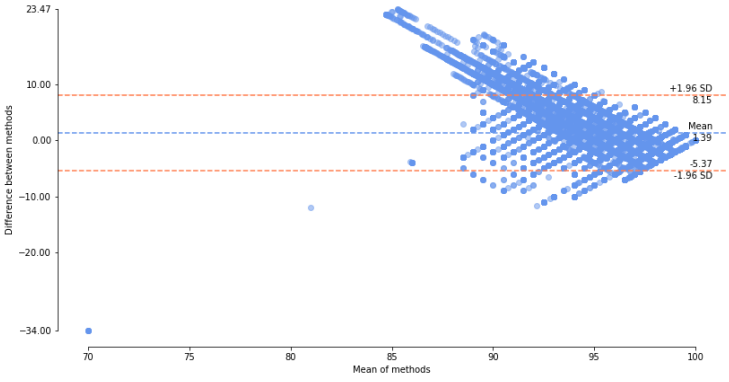
The results for each separate recording sorted from best to worst accuracy can be found in Appendix D while a summary calculation can be viewed in Table 7.10. The table gives the calculated metrics of all the signals in three different groups (all, recordings from lab, recordings from overnight monitoring). Then there is the mean and SD of the individually calculated metrics for the different groups. From this summary we see that the overall accuracy of all signals does not meet the ISO standards

requirement at 3.7%, or the FDA specification though not by much. The mean of the accuracy for all the data sets at approximately 3.01%±1.6 does meet the FDA specification. Of the 43 recordings (excluding outliers), approximately 67.4% of them had a mean accuracy lower than 3% and 79.1% was lower than 3.5%.

We have collected signals in two different ways; from a semi-controlled environment in a lab, and from unattended overnight sleep monitoring at the subjects home. These samples are not necessarily comparable, so we take a further look at the metrics for these subsets which is given in Table 7.10. Overnight is worse than the lab on all metrics, with accuracy at 3.48% compared to 2.84% for lab. The largest difference is with mean bias where overnight got 1.15 worse agreement score than lab. Despite the better mean performance of lab data, it also has a greater SD than overnight. The difference between the metrics were not statistically significant based on Wilcoxon rank sum test.



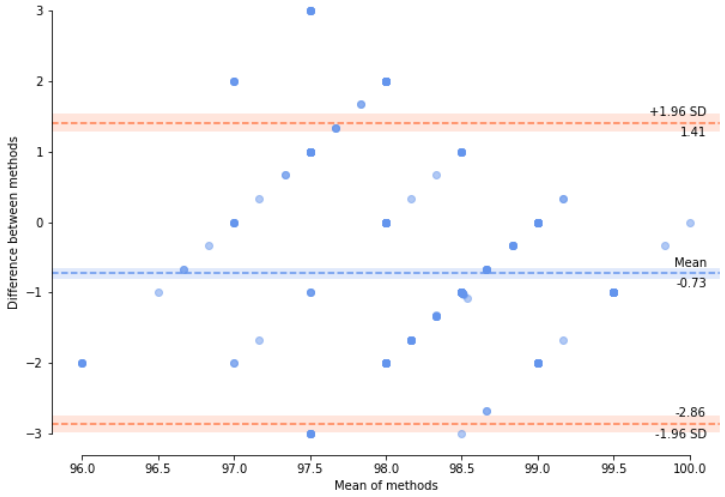
(a) *Scatter plot of signals from watch and Nox, r =0.16, p <0.0*



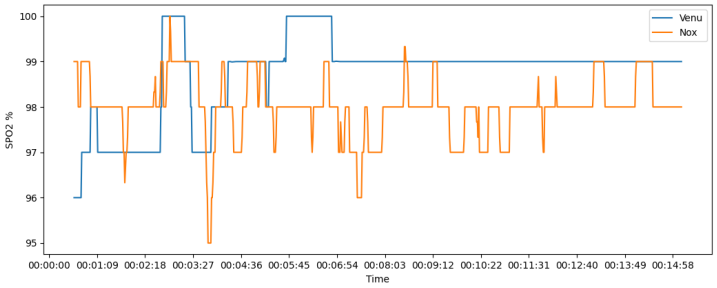
(b) *Bland-Altman plot of all recordings*

Figure 7.9: *Plots showing all data*

The best performance overall was by R-104 with accuracy of 1.31%. A closer look at the data however reveals that the signals are not quite that alike as can be seen in Figure 7.10b. Rather the low accuracy could be because of the low variation in the SpO_2 signals. This data set did not have any desaturation events, and for more than half of the duration the signal from Garmin read 99% oxygen saturation. Looking at the following four recordings with the best accuracy (Figure 7.13) we see that they also had little signal variation and did not quite match with the Nox signals.



(a) *Bland-Altman plot of R-104*

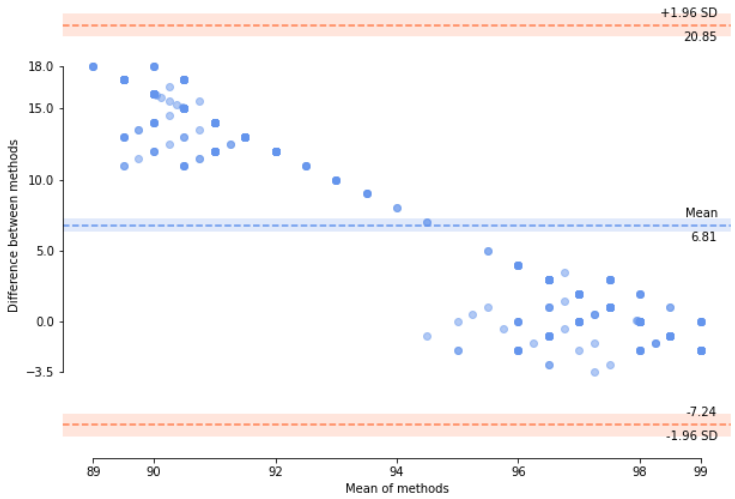


(b) *Graph of R-104*

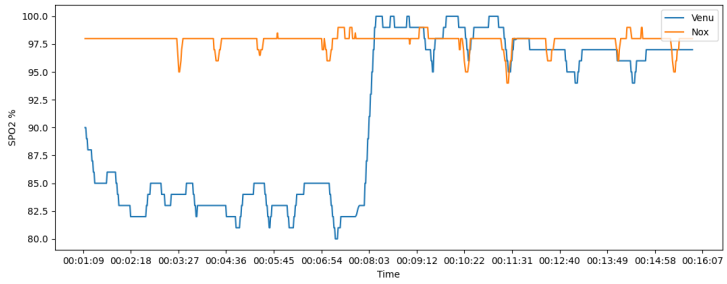
Figure 7.10: *Plot and graph of R-104*

The worst performance was R-112 with an accuracy of 9.88 and mean bias of 6.8. Looking at the graph of the recording makes it evident why the accuracy was so bad. In this experiment, the watch was worn tight. The SpO_2 levels increased from a lower level around the halfway mark, which is also about the time the subject switched positions from laying on their back to side. The plotted graph and Bland-Altman plot in Figure 7.11 reveals that the part of the recording where the signals do align has good agreement. It could be interpreted as the switch in position also impacted

the Venu oximeter's signal quality. This drastic contrast in SpO_2 values between positions was also seen in some other recordings.



(a) *Bland-Altman plot of R-112*



(b) *Graph of R-112*

Figure 7.11: *Plot and graph of R-112*

Of the following four recordings with the worst accuracy, of which their graphs are given in Figure 7.14, three of them are from overnight monitoring. There is large variations in the Garmin SpO_2 signals compared to Nox. It should be noted that Noxturnal processes the signals before we export them and we do not know what this processing is. At the worst for R-131 in the bottom left of Figure 7.14, the signals ranges between 75-100 for Garmin while only between 93-100 for Nox. R-123 in the top left is the last of the four recordings which is not from overnight monitoring. The graph shows a gradual decline in the oxygen saturation measured by Venu, starting at 97% and ending at 83%.

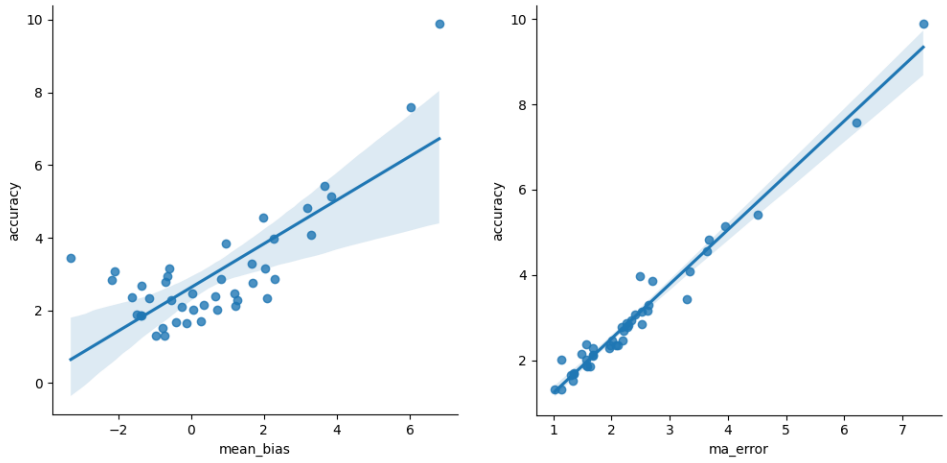
As some subjects performed multiple experiments, we also summarised quality metrics for each subject which can be seen in Table 7.11. Subjects 8, 11 and 13 only performed one experiment each which was the normal wear experiment. The metrics for these subjects are placed at the bottom.

The subjects are sorted from best to worst on accuracy. From this table we see that the best performing subjects are also the subjects that did not participate in the overnight monitoring (Sub007, Sub005, Sub004). Still, only four subjects had accuracy worse than 3%.

Metrics by Subject						
	Accuracy	**MAE**	**Mean Bias**	**Precision**	**Upper LoA**	**Lower LoA**
Sub007	1.95(±0.8)	1.63(±0.8)	0.31(±1.8)	2.80(±0.9)	3.05(±2.4)	-2.44(±1.3)
Sub005	2.23(±0.1)	1.72(±0.3)	0.81(±0.7)	4.03(±0.3)	4.76(±0.3)	-3.14(±1.0)
Sub004	2.34(±1.3)	1.81(±0.8)	-0.38(±1.2)	4.05(±3.0)	3.59(±4.1)	-4.34(±1.8)
Sub0015	2.57(±1.5)	2.08(±1.1)	0.19(±2.1)	4.11(±2.2)	4.22(±4.2)	-3.84(±0.4)
Sub009	2.58(±0.3)	2.02(±0.3)	-0.88(±0.4)	4.81(±0.5)	3.84(±0.7)	-5.59(±0.6)
Sub001	2.92(±1.1)	2.23(±0.9)	0.81(±1.7)	4.56(±1.7)	5.28(±3.0)	-3.66(±1.9)
Sub0014	2.95(±0.6)	2.59(±0.6)	-2.34(±0.9)	3.29(±1.3)	0.89(±2.0)	-5.56(±0.8)
Sub006	3.24(±1.3)	2.58(±1.0)	0.70(±1.3)	6.05(±2.2)	6.63(±3.0)	-5.23(±2.0)
Sub0012	3.68(±0.6)	2.99(±0.5)	2.48(±1.1)	5.25(±0.6)	7.62(±0.6)	-2.67(±1.7)
Sub003	3.99(±2.4)	3.26(±2.0)	1.93(±3.0)	5.91(±2.3)	7.72(±5.2)	-3.87(±1.2)
Sub002	4.51(±3.3)	3.48(±2.5)	2.81(±2.5)	6.95(±4.5)	9.62(±6.9)	-4.01(±2.0)
Sub008	1.70	1.35	0.28	3.36	3.57	-3.02
Sub0011	1.89	1.57	-1.49	2.32	0.79	-3.77
Sub0013	2.47	2.02	0.02	4.93	4.86	-4.81

Table 7.11: *Summary metrics grouped by subject, sorted from best to worst on accuracy*

We also plotted accuracy with mean bias and MAE to see their relation of which both can be seen in Figure 7.12. Both mean bias and MAE are strongly correlated with accuracy with r at 0.77 and 0.99 respectively. We therefore use accuracy as the main quality metric.



(a) *accuracy and mean bias - r=0.77, p≤0.00* **(b)** *accuracy and MAE - r=0.99, p≤0.00*

Figure 7.12: *Correlation between accuracy and metrics mean bias and MAE*

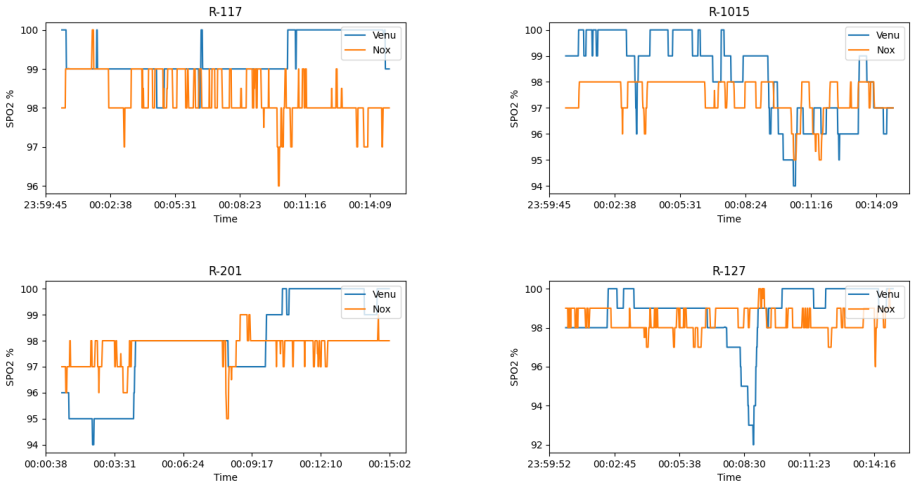


Figure 7.13: *Graphs of the best recordings*

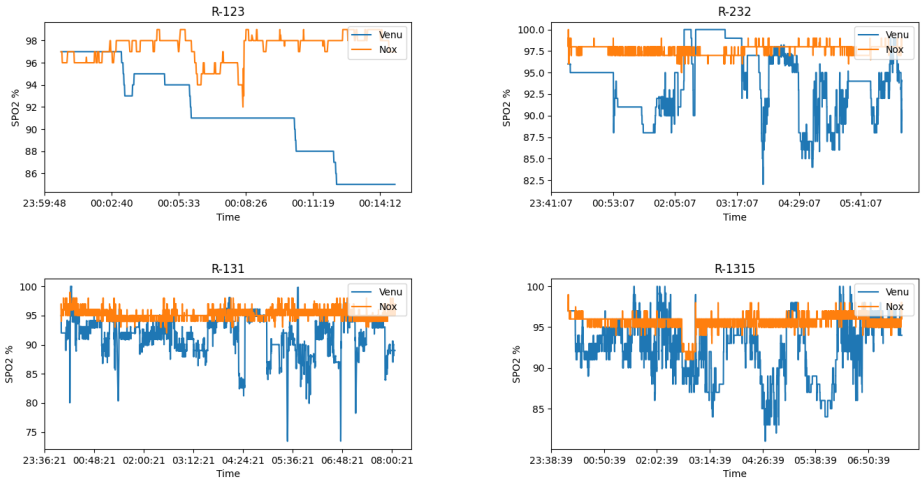


Figure 7.14: *Graphs of the worst recordings*

7.6.2 Factors Affecting Sensor Quality

The reflective method for blood oxygen measurement is known to be worse than the transmissive, and the pulse oximeters in smartwatches uses the reflective method. Watches additionally are placed in an area where there tends to be thicker skin and gets easily affected by movement. To better understand the results we will assess how the accuracy of the sensors are affected by external factors such as the way the watch is worn, skin tone and movement.

Despite the subjects following a breathing script for the experiments, there was not an even number of desaturations between the data sets. Some had multiple events while others had none. There was also individual differences in the number of desaturation events during the overnight monitoring. We will therefore also assess if the number of desaturation events has an impact on the Garmin watch's accuracy. We will assess the factors individually and also see if there is any interaction between the included variables.

Watch Wear

We performed experiments where we tested different ways the watch was worn on the wrist. Additionally, we performed overnight monitoring. We include overnight monitoring in the watch wear analysis, though one could argue against it since it does not have the same level of controlled setting. Our initial hypothesis is that wearing the watch tight or placing the sensor on the back of the wrist will give better accuracy than wearing it normal. The formalised null hypothesis (H_0) and the alternative hypothesis (H_A) is as follows:

$$H_0 : A_{rmsN} = A_{rmsT} = A_{rmsB} = A_{rmsO}$$

$$H_A : A_{rmsT}, A_{rmsB} > A_{rmsN}, A_{rmsO}$$

H_0 says that there is no difference on the accuracy between the groups, that they are all variations from the same population. H_A says that wearing the watch band tight or on the back of your wrist gives a better accuracy than wearing it normal. The reasoning behind H_A is that the reflective method for blood oxygen measurement is not effective on the wrist because of worse access to arteries. Wearing the watch tight removes some of the barrier between the sensor and the skin, while wearing the watch on the back of the wrist will place the sensor where the skin is thinner and have better access to arteries. As the watch is worn the normal way during overnight monitoring this is also included.

We grouped the metrics based on these criteria which can be seen in Table 7.12. The normal wear had the best overall performance, followed by back and tight with accuracy at 3.1 with overnight at 3.5. Already we see that H_A is false based on this table.

The overnight recording gives the most realistic performance of the watch. The results seen here are more consistent than tight and back based

Metrics by Experiment Type						
	Accuracy	**MAE**	**Mean Bias**	**Precision**	**Upper LoA**	**Lower LoA**
Normal	2.52(±1.0)	2.04(±0.8)	0.49(±1.6)	3.96(±1.6)	4.38(±2.6)	-3.39(±1.8)
Tight	3.10(±2.5)	2.45(±1.8)	-0.03(±2.7)	4.78(±3.6)	4.65(±6.1)	-4.71(±1.6)
Back	3.09(±1.9)	2.52(±1.6)	0.47(±2.6)	4.66(±2.4)	5.04(±4.7)	-4.10(±1.4)
Overnight	3.48(±1.2)	2.69(±1.0)	1.48(±1.7)	5.72(±1.4)	7.08(±2.8)	-4.13(±1.2)

Table 7.12: *Mean and SD of signal quality metrics for each experiment variation*

on the SDs, however the mean of the metrics are worse than the scripted experiments. The normal group had a mean average below 3% (2.52%±1) which means it meets the ISO standards requirement. There is complete overlap of the accuracy from the SD's for the different groups. This is also reflected in the ANOVA test in Table 7.13 not being significant. Based on these results we cannot reject H_0 that the mean accuracy of the groups are from the same population.

ANOVA					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
experiment	3	5.961	1.987	0.728	0.542
Residual	39	106.511	2.731		

Table 7.13: *ANOVA test of different experiment groups*

Skin type

Figure 7.15: *Fitzpatrick skin types [18]*

Our hypothesis is that the darker the pigmentation the worse the quality of the sensor. This is based on previous research on the topic [6, 61]. The skin types are based on the Fitzpatrick classification which consists of the six types given in Figure 7.15 [28]. The Fitzpatrick scale is commonly used for identifying different skins reaction to UV light. In our case, we use the descriptions of the types and not the characteristics that are assessed. We compress these into the three categories of light (Type I, II), medium

(Type III, IV) and dark (Type V,VI). A null hypothesis and an alternative hypothesis have been formalized as follows:

$$H_0 : A_{rmsL} = A_{rmsM} = A_{rmsD}$$
$$H_A : A_{rmsL} > A_{rmsM} > A_{rmsD}$$

H_0 states that the three groups are from the same population, meaning the difference in accuracy is due to random sampling. H_A states that the groups are different and that lighter skin tones have better accuracy than darker skin tones.

Metrics by Skin Type						
	Accuracy	**MAE**	**Mean Bias**	**Precision**	**Upper LoA**	**Lower LoA**
Light	2.62(±1.0)	2.16(±0.8)	0.09(±1.8)	3.88(±1.5)	3.89(±3.1)	-3.72(±1.4)
Medium	2.63(±0.9)	2.09(±0.7)	-0.22(±1.2)	4.78(±1.8)	4.47(±2.6)	-4.90(±1.5)
Dark	3.44(±2.1)	2.68(±1.6)	1.35(±2.3)	5.35(±2.9)	6.60(±4.9)	-3.89(±1.7)

Table 7.14: *Metrics based on skin tone*

Sorted by skin tone we have 27 hrs, 6 min for light, 16 hrs, 26 min for medium and 37 hrs, 51 min for dark. A summary of the metrics grouped for different skin tones can be viewed in Table 7.14. From the summary we see some variations between the three different groups. The first observation is that the accuracy, MAE, precision and mean bias is worse for the dark group, while light and medium is pretty much the same on accuracy but not precision and mean bias. There is also higher variation in the dark group compared to light and medium. From the SD of the accuracy means we see there is complete overlap for the three groups. The p-value from the ANOVA test is 0.265 which is not significant, though the low p-value can be attributed to the higher mean for dark. We therefore based on this cannot reject H_0 that the means are from the same population.

ANOVA					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
skin_type	2	7.221	3.611	1.372	0.265
Residual	40	105.251	2.631		

Table 7.15: *ANOVA test with interaction*

Interaction

We have analysed how the two variables "watch wear/experiment type" and "skin tone" affect signal quality separately, however there might be some interaction between these variables. The reasoning behind this is that overnight and dark got the worst mean accuracy, and the dark group also had the largest number of overnight samples. To evaluate this further we perform an ANOVA test for the variables skin type and watch wear.

ANOVA					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
skin_type	2	3.025	1.513	0.536	0.590
experiment	3	5.544	1.848	0.655	0.586
skin_type:experiment	6	13.171	2.195	0.778	0.593
Residual	32	90.315	2.822		

Table 7.16: *ANOVA test with interaction*

From the ANOVA table in Table 7.16 we see that none of the individual predictors are significant, and neither is the interaction. Despite this, we see that the p-value for skin type got worse ($p = 0.59$) when both experiment and the interaction was included. This can be interpreted as when adjusted for experiment type, the effect of skin type disappears.

Movement

Our hypothesis is that more movement leads to worse sensor quality, while the null hypothesis is that movement has no impact on the data. This is based on that more movement will also lead to the Garmin watch moving more.

$$H_0 : \textit{movement does not impact accuracy}$$

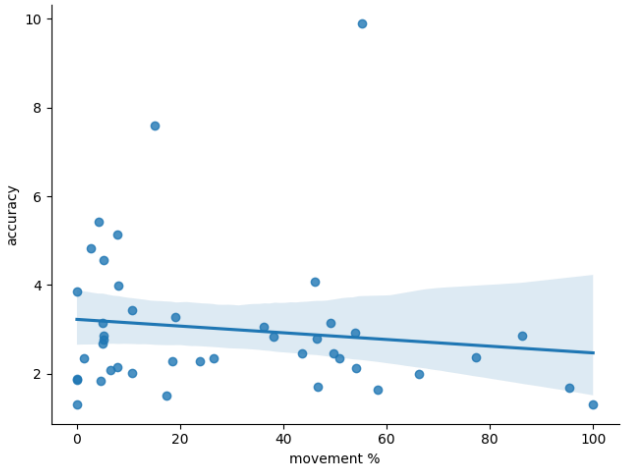
$$H_A : \textit{more movement leads to worse accuracy}$$

Nox T3 registers movement as events from the RIP bands. The number of movement events for each recording will be counted and divided on the total length of the recording and multiplied by 100 to get the percentage of movement in the data set. These values will then be plotted against the accuracy in a scatter plot and a regression line will be fitted with a 95% confidence interval.

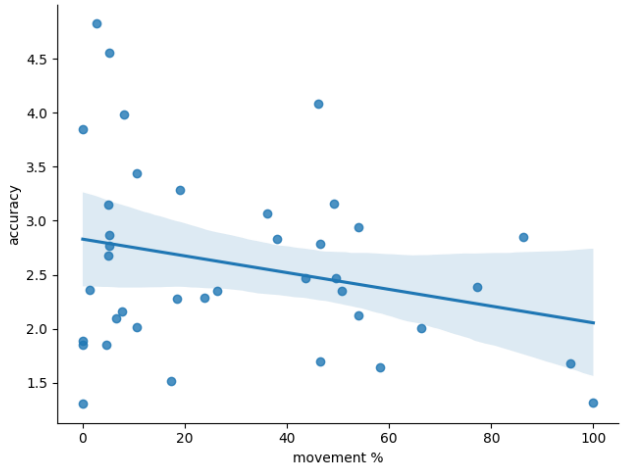
In Figure 7.16a all recordings are plotted. From the scatter plot there seems to be some extreme values that are influencing the regression line. We therefore use inter-quartile range to calculate upper and lower bounds to uncover any outliers. For accuracy the lower bound is 0.308 while the upper bound is 4.968. Based on this there seems to be four outliers in the plot which are R-112 (9.883), R-123 (7.582), R-232 (5.416) and R-131 (5.135). The scatter plot without the outliers can be seen in Figure 7.16b.

There seems to be a weak correlation between accuracy and movement, both with (r=-0.13) and without (r=-0.26) outliers for accuracy. Furthermore the correlation is negative, meaning the accuracy gets better with more movement. This is the opposite of our hypothesis. However the correlation is not significant, which means we do not reject H_0 of movement not impacting signal accuracy.

The reliability and validity of the registered movement events in the Noxturnal software should be questioned. The signal capture procedure



(a) *with outliers - r=-0.13, p=0.40*



(b) *without outliers - r=-0.26, p=0.10*

Figure 7.16: *Correlation between accuracy and movement (%)*

required the subjects to move from one position (supine) to another (back), which means there should be some movement for all lab tests. This was however not the case as some recordings registered no movement events. At the same time, there was a high level of movement for some other recordings despite the subject lying still for most of the time. In the future another external form of assessing level of movement should be considered, or the accuracy of Noxturnal's labelling should be tested.

Desaturation Events

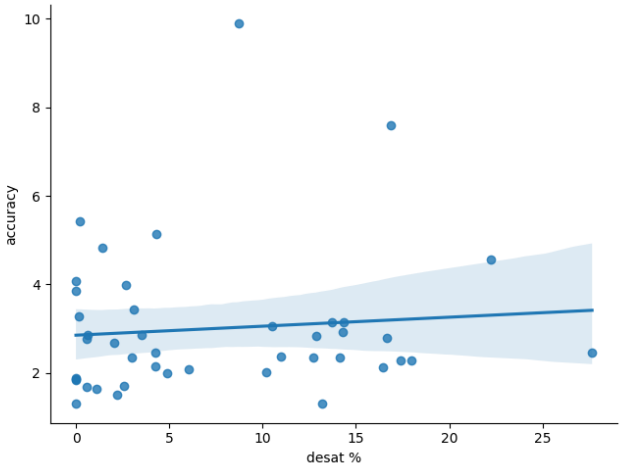
The null hypothesis is that the number of desaturation events does not affect the sensor's accuracy, while the alternative hypothesis is that more desaturation events worsens the sensor's accuracy.

$$H_0 : \textit{desaturation does not impact accuracy}$$

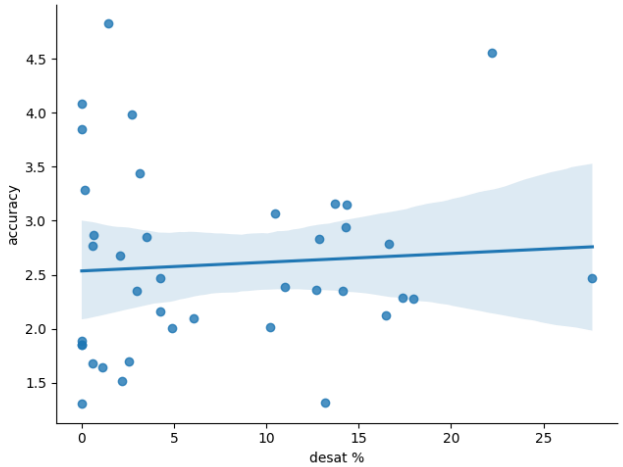
H_A : *more desaturation leads to worse accuracy*

The reasoning behind this hypothesis is that Garmin watches might perform well under normal circumstances. With high level of desaturation events in the data the Garmin watch might not be as effective. As is already known, the accuracy for pulse oximeters varies at different ranges according to the ISO-standard. In our case, we have seen from the histogram of recorded signals that the majority of the signals falls in the range 70-100%. The accuracy therefore should be $\leq 3\%$.

We also witnessed differences in the number of samples for SpO_2 values in the histogram in Figure 7.7c. The majority of Nox's values was between 90% and 100% and some between 80% and 90%. On the other hand, the watch registered values between 75% and 100% with a fuller left tail. In this case we hope that the null hypothesis is true because it supports our overall goal of using the watches for at-home detection.



(a) *with outliers - r=0.09, p=0.56*



(b) *without outliers - r=0.07, p=0.67*

Figure 7.17: *Correlation between accuracy and desaturation event (%)*

We use the same approach as with movement which is calculating the percentage of events in the recording. Figure 7.17a shows the scatter plot with outliers while Figure 7.17b shows the scatter plot without. Both plots show a weak positive correlation (with: r = -0.09, without: r = -0.07) which are not significant (with: p = -0.56, without: p = -0.67). We can therefore not reject H_0 of desaturation does not impact signal accuracy.

7.6.3 Comparing With Related Work

There has been some studies evaluating the quality of consumer pulse oximeters to medical grade oximeters by Frisvold[19] and Harskamp et al. [27], some also on the pulse oximeter in smartwatches by Halvorsen [26] and Lauterbach et al. [38]. Frisvold's study evaluated the quality of the consumer pulse oximeter Cooking Hacks MySignals with Nox T3 as reference. The results were much better than ours with a mean accuracy at 1.34% for all signals and a mean bias at 0.14% ($\pm$2.61). The worst accuracy was 2.09%. Harskamp's study got more similar results to ours. 10 pulse oximeters were tested and none of them met the ISO requirement of accuracy $\leq$ 3%, the lowest being 3.9% and highest 7.5%. Mean bias ranged from -0.6 to -4.8. Halvorsen's results are directly comparable to ours as they use both the same brand of watches and reference sensors. The accuracy ranged from 1.6% at best to 8.2% at worst. Lauterbach assessed Garmin Fenix 5X Plus pulse oximeter to a reference oximeter at five different altitudes. A Bland-Altman analysis was performed an mean bias ranged from 0.0 at the lowest altitude to 3.3 at the highest. From these four studies we see that there is a large variability in the results.

7.6.4 Summary

With the signal quality evaluation we aim to assess whether the SpO_2 signals from Garmin were as good as Nox. The results can be summarized as follows:

- Almost 70% of the individual data sets meet the requirement of < 3% accuracy, while almost 80% where under FDA's typical specification at 3.5% for reflectance pulse oximeters. The average mean bias for all data sets was 0.62%.
- The results were slightly worse for overnight monitoring (accuracy: 3.48($\pm$1.2), MAE: 2.69($\pm$1.0), mean bias: 1.48($\pm$1.7)) compared to lab (accuracy: 2.84($\pm$1.8), MAE: 2.29($\pm$1.4), mean bias: 0.33($\pm$2.2)) on all metrics.
- The worst performing subset of data is from overnight monitoring, which is the most realistic. However, it also had one of the smaller SD compared to other subsets.
- All four hypotheses regarding how variables (skin type, watch wear, movement, desaturation event) impact signal quality were rejected.

This could be because of the small size of the samples. They should still be investigate further because of observed trends in the data.

- There were still some interesting trends such as normal watch wear performing the best, more movement being correlated to better accuracy and dark skin type performing worse than medium and light.
- The level of movement in a data set was determined by the Noxturnal software's scoring. The scoring did not seem to be accurate as not all lab recordings registered movement even though all subjects moved from lying on their back to their side. In the future, the reliability of the scoring should be tested our an alternative way of assessing movement should be used.
- Our results are comparable to previous work in regards to accuracy and mean bias.

Chapter 8

Discussion

The results we obtained will be further discussed in this chapter. We will try to understand and make sense of the results we got by discussing the results for the different tests we performed. For instance, our results from the connection loss detection differed from previous tests. We try to make sense of this in (Section 8.2). We also discuss the performance of our classifiers (Section 8.3), the quality of our data (Section 8.4), and discuss how they could have impacted the classification performance (Section 8.5).

8.1 Usability testing

For usability testing we had three participants perform two tasks in the app. They vocalized their actions while the test manager observed their performance. Questions were asked both before and after the tasks.

Of the three participants, one of them was not as representative of the target audience as initially thought. That the particular participant, nr. 3, was not representative was discovered during the introductory questions. Participant 3 mentioned that they were not that interested in using smartwatches, and is not an avid user of apps on their phone. The results are still included as the participant represents an inexperienced user. Despite this the participant managed to complete both tasks and did have some relevant insight to the app. In the future, recruiting of participants should include a more in-depth initial screening.

There seemed to be more errors in this round of usability testing than the one performed by Halvorsen, though not by much. All three participants in this round mistakenly passed the clicking of the device name before showing paired devices. This error and feedback led to the adding of a descriptive text over the listed devices. Furthermore, clicking the button when there are no paired devices will display an error message. We see that taking a more qualitative approach to the usability testing led to more concrete solutions.

The app was unstable during testing with Participant 2 and 3. It crashed during the pairing of device. This was however not fixed due to restrictions in time. If this app were to be developed past the Minimum Viable Product (MVP) level, then this is something that should be fixed, along with the

other feedback on the app.

8.2 Differing Results for Connection Loss Detection

Both the experiments with short duration and the long duration were repeated experiments originally performed by Halvorsen. From Halvorsen's results there was a shorter disconnect registered in the data than the planned disconnect duration of one minute and one hour, indicating data being buffered. These results could not be reproduced. For both experiments where we experienced disconnect, they lasted for the exact length specified or longer. The reason for it being longer might be due to the time it took to get back within Bluetooth range. There is a possible explanation that the updates that were made to the Companion SDK might have affected this. This theory has however not been researched so it cannot be supported.

We also performed a second experiment for short duration where we tested Garmin's Bluetooth range. Garmin states the range for their Bluetooth is 13 meters without barriers. In our test the distance was approximately 15 meters with a wall in between. We still did not experience any disconnect. This is a positive result as it means a scenario where the user wakes up in the middle of the night and moves around their house without their phone might not necessarily lead to connection loss.

The last test of close proximity with barrier also resulted in no connection loss. This means the connection is strong enough to withstand the user moving around throughout the night.

There is a clear weakness in the connection loss test which is we cannot accurately determine the exact moment the device is outside of communication range.

8.3 Classifying Desaturation Events

For desaturation event classification we tested four different classifiers on SpO_2 signals from two different devices. For holdout testing of each individual recording we used all the lab data for training the model where the test set was a lab recording, and when the test set was an overnight recording we trained the model on all the overnight data. For the ODI classifier we labelled each second before we reshape the data to 60-second periods.

An immediate issue with the ODI classifier is that it requires 110 signals before the first event can be detected. That is, the first desaturation event can at earliest be detected at signal 111. This is an issue because events that occur before this will never be detected. From our evaluation we see that the classifier got an accuracy a high as 0.91 and there was not much difference in performance between devices which is different from the ML classifiers. However, the highest κ was 0.15 and the sensitivity was nonexistent while specificity was at around 1.0. Since it is an outdated

classification method and the performance was so poor compared to the ML classifiers, we do not consider this method any further.

From the holdout testing of individual recordings we see that the results are not that different compared to 10-fold CV. A factor that does seem to largely impact performance is small data sets. The lab experiments on the holdout test performed worse than overnight test for all classifiers. The windows were 60 seconds, which might have been too large for such small data sets. For future testing, overnight recordings could be included for training the models, or use a smaller window size for smaller data sets e.g. 30 seconds. Another possibility is to train the models on data with better quality i.e. training the models on data from Nox then testing on Garmin data. This has been done in an unpublished article by Kristiansen et al. [33]. In the study, the accuracy of classifying low-quality data from a sensor called Flow was nearly as high when training on high-quality Nox data as the low-quality Flow data.

Because of our sample sizes being so small, we only used simple ML classifiers. In the previous work that have been compared in this thesis, more advanced algorithms were used such as Neural Network (NN). These produced even better results, particularly Convolutional NN.

8.3.1 Garmin vs. Nox

It was clear from the means of the different classifiers that the performance on Nox data was far better than on Garmin data. This was also significant from Wilcoxon rank sum test. The κ and accuracy obtained from the classifiers on Nox were comparable to previous work with κ at 0.4-0.5 and accuracy at 0.6-0.8 for all the data despite the small sample. As both devices sampled the same signals, we attribute the difference in performance to the quality of the sensors. With this we mean the Garmin pulse oximeter has poorer quality than the Nox T3. There is however not a direct comparison between the two devices signals, since the signals sampled by Nox is processed by algorithms in Noxturnal. Such processing could be removal of noise of some sorts which we do not know.

8.4 Sensor Quality of Garmin Venu 2S

Comparing the quality results with Halvorsen's we see the results are fairly equal in regards to accuracy. The Venu performed better than Fenix in his experiment while the opposite was true for the overnight test. For both watches the best accuracy was 1.6% while the worst was 8.2%. In comparison, for our experiments the best accuracy was 1.3% and the worst was 9.9%. Both studies observed large variations in regards to performance.

Frisvold also performed oximeter quality test, though not with a smartwatch. The accuracy of the experiments were consistently between 0.5 to 1.5% with the exception of one at 2.09%. In Harskamp's study of ten consumer pulse oximeters, none of the oximeters got an accuracy smaller

than 3%, ranging from 3.9-7.5%. From these results it is clear that though the Garmin sensor can perform at the same quality as other low-cost pulse oximeters, it is less stable and more prone to errors.

8.4.1 External Variables Impact on Signal Accuracy

We tested if the four variables movement, desaturation, watch wear and skin type can affect Garmin signal accuracy. Even though none of the tests were significant, we still observed some trends in the results. From the test on skin pigmentation we saw that darker skin had worse accuracy and larger variation on all metrics. Similarly, overnight monitoring had worse accuracy than the lab recordings. As the results could have been affected by the fact that the dark group had a larger number of overnight monitoring included compared to the other two groups, we tested the interaction between watch wear and skin type. We also see the same tendency of subjects who performed overnight monitoring also had the worst metrics. Though the results of the interaction was not significant, we observed that the p-value for skin type was worse when experiment and their interaction were included in the ANOVA model.

An interesting result was that wearing the watch normal had the best performance of the three ways of wearing the watch. This was unexpected as measurement site affects accuracy (transmittance compared to reflectance), and wearing the watch tight or placing the sensor directly on arteries would seem to improve accuracy. Due to the small sample size and there not being any significant difference we cannot draw any definite conclusions. This should, however, be tested further as it supports the quality of pulse oximeters in smartwatches. In the future, this could be assessed in overnight monitoring as well.

An observation we made was that the placement of the arm sometimes affected the signal quality. For certain recordings we saw a difference in signals from when the subject was lying on their back or side. That is, for some recordings the SpO_2 levels could be around 85% when the subject was on their back, but after switching positions it could be at 95%. We believe this is because of how the watch is laying on the arm in different positions. Due to time constraints, this was not investigated further. In the future, this should be assessed.

In our tests we only assessed the impact of four different variables on accuracy. In reality there are many more variables, known and unknown, that can affect the signal quality. From what we observed during the lab tests, the position of the subject affects the signal. Most likely because it affects the watches placement on the wrist. Also, the subjects varied in the amount of hair on the wrist under the sensor. The watch fit some subjects better than others as well. If possible, such variables should be controlled for in future studies. Even though the variables we tested did not significantly affect signal accuracy, there was some results that could be explored further.

8.4.2 Removal of Outliers

Some of the outliers were obvious that should be removed such as the total duration being too short or oxygen saturation greater than 100%. These were removed during preprocessing and not included in any of the results. On the other hand, there were outliers that were extreme values that deviate from the majority, like signals at around 60% or recordings with accuracy at 9%. These were discovered upon further examination of the results. Such values strongly impact the results. The question about whether such outliers should be included or removed comes down to whether the extreme values represent possible and realistic observations.

Upon further inspection we observe that most of the data sets with worse accuracy were also the data sets from overnight monitoring. In regards to the lab recordings it was the data sets where there was large deviations between signals when the subject lied on their back compared to on their side. For the correlation studies where we tested different hypotheses, the outliers were removed to see if there was any difference in the results. Scatter plots with and without the outliers were placed side by side to visualize the difference. The differences were minimal.

8.4.3 Metrics

A total of three different metrics were used to assess signal quality. These were Accuracy root mean square (A_{rms}) which is the standard used for oximeter quality, Mean absolute error (MAE) and mean bias by Bland-Altman. There has been some discussion against the use of Root mean square error (RMSE) which is the same formula asA_{rms}. We plotted these measures against each other and got $r = 0.99$ and mean bias $= 0.61$. Accuracy and mean bias were also strongly correlated with $r = 0.77$. The decision of metrics to use falls on preference and what we are trying to tell with our data.

8.5 Overall Data Quality

The majority of the collected data sets were from the breathing script. These were of a shorter duration, and was therefore easier to perform as one can do multiple recordings in a day with awake subjects. This is, at the same time, also a drawback as it does not reflect the natural setting. Even though we managed to perform many more recordings with the script, the total duration was significantly shorter at approx. seven hours compared to overnight monitoring at approx. 74 hours.

The benefits of the breathing script as data collection method for quality estimation is that we can control the number of events (to a certain degree), and the shorter length of each data set makes it easier to analyse. Unfortunately, there were large differences in the number of simulated events, as some subjects got none while others got up to five events. All subjects followed the same breathing script, they tested the breathing styles beforehand and no one struggled with them. There was

individual differences as, for instance, Subject 7 and 4 held their breath for as long as they could but Nox T3 still did not register a desaturation event. Nevertheless, the lab tests had a higher percentage of events than overnight at 32.5% compared to 7.7%. The quality on overnight monitoring was worse than for the lab recordings which might be due to the controlled nature of the latter. However, there was less variability in the results for overnight monitoring.

For classification, we saw that having more data resulted in better performance for the 10-fold CV. When tested on the separate watch wear lab tests only, the highest κ was 0.27 (accuracy = 0.6) for SVM on back experiment for signals from Nox. In contrast, the lowest κ for overnight monitoring was 0.49 (accuracy = 0.74) by KNN. This trend was not present for Garmin data where all κ ranged between -0.2-0.2 (accuracy = 0.3-0.6). From this we see that larger data quantity does not matter if the quality is not good. Additionally, consistency of the accuracy in the data set seem to matter. The overnight data set performed slightly better than all the data, even though the latter is larger. This could be attributed to the larger variance in accuracy of the lab data, which we saw from the quality assessment.

To summarize, there were trade-offs for both methods of data collection. Classification benefited from the quantity from overnight monitoring, while quality was also worse for this data. Still, the preferred data collection method would be through overnight monitoring as it has direct comparability to the intended use. For better results and the possibility to draw any actual conclusions, there needs to be a larger database of good quality and from a representative population such as subjects with SA or other respiratory disorders to test on. The results are promising for the Garmin pulse oximeter quality. However, it should be noted that the Garmin signals were tested against another pulse oximeter. If they were to be tested against a CO-oximeter, which is the recommended method for accuracy assessment, the results would be worse. In regards to classification, there is some uncertainty due to the low amount of data. The results are less promising but not discouraging.

Part IV

Conclusion

Chapter 9

Summary of Contributions

In our introduction we define four questions that we want to investigate in order to answer our problem statement of: *Can a Garmin smartwatch's pulse oximeter be used for initial at-home sleep apnea detection?* With these questions in mind, we performed several test. In this chapter, we summarize the contributions of our evaluation. We tested the usability of the app created for collecting data from Garmin sensors. Also, we tested the boundaries for Bluetooth connection between a smartwatch and the Garmin watch and connection loss detection with both short and long duration (Section 9.1). Since there was no already existing signal data from both Nox T3 and Garmin which were comparable, we also recorded the data to be analysed (Section 9.2). For classification, we used four classifiers for labelling on both Garmin data and Nox data (Section 9.3). We also evaluated the quality of the Garmin signals with Nox as the reference pulse oximeter. Then, we evaluated how four different variables affect the accuracy (Section 9.4).

9.1 Reproducing Previous Results

Of the several tests performed by Halvorsen, we repeated the usability test and connection loss detection for the app, and classification with a signal counting script and quality assessment for recorded signals. We will explore more in depth about classification and quality in the following subsection as we did many things differently. The purpose of repeating these tests, specifically the app tests, was to evaluate reproducibility of the results and evaluate the potential in the app with Garmin Health Companion SDK.

The two main functions in the Cesar smartwatches app are to pair the phone with a Garmin smartwatch and start and stop recording signals from sensors. These two functions were tested in the usability test. We uncovered some ambiguity in the app as all participants mistakenly overlooked clicking the device name for pairing, and there was confusion of whether the watch was paired or not. Performing the usability test allowed us to better understand some issues and therefore also fix them. We updated the app to include text over the listed devices, the "Show paired"-button is no longer clickable if there is no already paired devices,

there is a progress-bar for when the pairing is happening, and informative toast messages are displayed so the user is more aware of what is happening. An issue that was not fixed was the app crashing during pairing. This should be fixed if the app is to be used as more than a MVP. Other than this, the participants found the app simple and easy to use.

With connection loss detection we tested the boundaries of the Bluetooth connection between the app and Garmin watch, and the reconnection implemented in the Garmin SDK. We found that the specified maximum range of 10 meters was inaccurate for our case. At 13 meters with walls in between there was still no disconnect. For the tests where we experienced disconnect there was no buffered data as was seen in Halvorsen's tests. The disconnection lasted for as long as the specified time. We also tested disconnect in close proximity but with barriers. This did not lead to connection loss. From these results we see that the Bluetooth range is better than initially specified, but there is no buffered data when the devices disconnect.

The first research question goes as follows: *Can the results from Halvorsen's experiments be reproduced?* We see from this summary that the usability test resulted in the app being simple and easy to use, same as Halvorsen's results. We were not able to reproduce the buffered data from the connection loss tests.

9.2 Data Collection

There are existing databases of SpO_2 data from medical grade devices which have been used in previous research. The problem is, however, that we want to compare the Nox T3 oximeter data with Garmin data, which does not have an existing database that we know of. We therefore had to record new data with Nox T3 and Garmin Venu 2S simultaneously. From 15 subjects we collected a total of 46 data sets from each device, of which 43 were used in evaluation. The collected data sets are from two different methods, in a lab with a breathing script/procedure lasting around 16 minutes and from unattended sleep monitoring at home. This gives us an opportunity to compare the two quality of the two methods. Additionally, since we use subjects without any SA diagnosis, simulating events in a lab can give us more representative data. We got on average 32.5% apneic windows from the lab tests compared to 7.7% from the sleep monitoring. With the lab recordings we also record with the watch worn in three different ways which is normal, tight and with the watch face on the back of the wrist.

9.3 Event Classification

We trained and tested three different ML models in addition to a signal count classifier we nicknamed the ODI classifier. For the ML classifiers KNN, SVM and RF, we used two different evaluation methods, holdout test set and 10-fold CV.

From both the holdout tests and CV, we see on all metrics that the performance was better on Nox data than Garmin data, with the exception of the ODI classifier. κ on Nox data was at its best 0.51 from holdout tests (RF on overnight), while the best from CV was 0.57 (RF on overnight). In comparison, the best κ on Garmin data was 0.32 (KNN on back), while for CV it was 0.15 (RF on all data). A possible solution to improving the performance on Garmin data is by training the model on better quality data. We observed that the RF classifier performed the best of the ML classifiers. Due to the small sample sizes, which proved to be detrimental to the performance based on the CV results, we do not draw any definite conclusions.

Another research question is the following: *Can using Machine Learning for classification improve event detection in Garmin signals?* From the results we conclude that using ML for classification does improve results overall, as sensitivity and specificity was more balanced compared to ODI where specificity was 1.0 and sensitivity was 0.1. Even though κ and accuracy was at times worse or the same for the ML classifiers as the ODI classifier for Garmin data, there is potential in fine-tuning of the models, or using a different model altogether.

9.4 Quality Assessment

We assessed the quality of the Garmin signals with the metrics accuracy (A_{rms}), Mean absolute error (MAE), and mean bias with Limits of Agreement (LoA). The accuracy of the data sets ranged from 1.305% to 9.883% with the mean at 3.01%($\pm$1.6). Nearly 70% were under 3%, which is the ISO-standard for oximeter accuracy. Furthermore, the accuracy was worse for the overnight data (3.48($\pm$1.2)) compared to lab data (2.84($\pm$1.8)). Even though this is unfortunate since the overnight data most accurately represents the real setting, we see that there is less variability in the overnight data. With more data we can get a more accurate estimate of the accuracy of Garmin's pulse oximeter.

In regards to sensor quality, we had the following research question: *Are the Garmin oximeters quality good enough for at-home OSA detection?* Based on the ISO-standard of $\leq$ 3% and the fact that we tested against a different pulse oximeter and not a CO-oximeter, we cannot say yes. However, with a mean at 3.48($\pm$1.2) for overnight and 3.01%($\pm$1.6) for all data it is not that far off. As it would only be used for initial detection, we would say this is acceptable if the results were more consistent. Also, it requires that we can accurately label events in the data, which requires a lot more data and work.

We tested whether the four variables watch wear/experiment type, skin type, movement, and desaturation is associated with accuracy. Although none of the hypotheses produced any conclusive results, we saw some trends that should be highlighted. For one, wearing the watch in the normal way which is not too tight does not negatively affect the signal accuracy. Darker skin type had worse accuracy and mean bias compared

to light and medium skin type, and had more variability on the metrics measured by higher SD. Movement was actually negatively correlated with accuracy, which means the more movement in the data lead to better accuracy. We used movement as it was labeled in the Noxturnal software, and later saw that it did not reflect our own observations. For this reason, we do not consider this result as valid. Desaturation as labeled by Noxturnal in the data was not significantly associated with worse accuracy.

The last research question is the following: *What external factors affects the quality of Garmin's pulse oximeter and how?* Based on the four variables tested, we can say that experiment type, or more accurately, data collection method (lab, overnight) and skin type affects Garmin signal accuracy. Both variables affected accuracy even though they were not significant. The results are still important as the lab vs. overnight monitoring sheds light on how the results in a controlled setting differs from a normal setting, while the skin type shows how quality may vary based on individual differences.

Chapter 10

Open Problems

Through our research we discovered some limitations in our methods, or areas that could be improved. Because of the restricted time frame, some of these were not directly addressed. For a large part of the project there were still restrictions in place due to the pandemic, which prevented us from collecting data from a planned group of subjects. Because of this, we were not able to recruit participants that actually have breathing disorders. This heavily affected the sample size and quality of our data, and also played a part in our choice of traditional classifiers.

From the data sets we saw in some of them that the signal quality differed from when the subject was on their back compared to on their side. We did not however evaluate what impact this had on the accuracy. The reason for this is that we could not clearly define the time the subject switched positions in the data.

We hypothesised that more movement leads to worse signal accuracy. However, we saw that the event labelling of movement as it was done in Noxturnal did not accurately reflect what we ourselves observed. Furthermore, the accelerometer is located in the main unit placed on the subjects chest. This will record different movements from the Garmin watch's accelerometer on the subjects wrist. Because of lack of time, we did not look into this any further. In the future, movements should be classified by the watches accelerometer, or observed externally during the lab tests.

Our method for synchronization is still not optimal as it relies too much on visual inspection of the graphs. Using the accelerometer for synchronization is not valid as the watch's sensor is placed on the wrist while for the Nox it is on the chest. In the future, a more reliable and valid method should be used in conjunction with they delay finding in the oxygen data.

Our ODI classifier estimates the baseline as the mean of the oxygen saturation of stable breathing in the two minutes preceding an event. This assumes that the preceding two minutes are stable, which is not always the case. In the occasion when the breathing pattern is not stable, the AASM suggest the use of the mean amplitude of the three largest breaths in the two preceding minutes instead. Our classifier has not implemented this as

we cannot detect breath amplitude in oxygen saturation data.

Chapter 11

Future Work

There are many opportunities in future work related to investigating performance of Garmin smartwatches' pulse oximeters, in regards to quality and classification. The main question is, despite our varied results, is the use of Garmin watches still promising? If the answer is yes, then some of the future work to be done for classification is presented in Section 11.1 and for signal quality in Section 11.2.

11.1 Classification

The biggest drawback in our evaluation was the lack of data. In the future, more data needs to be collected, preferably from overnight sleep monitoring from SA subjects. If this is done, then we can use more advanced algorithms for classification, and also there will be less need for balancing the data sets. Another opportunity to improve results would be by training the model on better quality data than the test set. That is, we could train the model on data from Nox and test it on data from Garmin. Data from the A3 study that has been used in the CESAR project could be used for training the model for testing on Garmin data.

For event detection we focused mainly on scoring/labeling events in fixed 60-second periods in the time series data. The periods in reality are not this fixed, they might vary in length, start and end time. In the future, the focus should be on finding the beginning and end of each event period. In a recent thesis by Guðmundur Jónsson [32] this was investigated. With this approach, we can also more accurately count the number of events per hour according to the Oxygen Desaturation Index metric. Another option for classification could be to classify severity based on comparing different time series data with each other.

11.2 Signal Quality

There is potential in the quality of the Garmin pulse oximeter, but more testing is needed. Furthermore, our analysis compares the raw data from Garmin with processed data from Nox T3 which does not make the

comparison equal. Nonin also describes the removal of noise artifacts from their oximeter signals to improve accuracy [52]. In the future, a similar approach should be taken with the Garmin signals. That being removal of noise and other forms of preprocessing of the signals.

What is most interesting is how different variables affect signal accuracy, and should be tested further. In addition to the variables tested in this thesis, future work should also include other variables such as temperature and position of the arm. This should also be tested in overnight sleep monitoring to see if the variables have an effect in a natural setting. An important factor to consider is that even if the Garmin oximeter's quality is up to standard, it does not matter if the data is not good enough for classifiers to make sufficiently good predictions.

11.3 App

If the plan is to move forward with Garmin, then data analysis should be added to the app. The CESAR project currently only has a research license for the Garmin Health Companion SDK. They will need to acquire a commercial license if they were to continue collecting Garmin signals. Integration of this app with the Nidra app previously created in the CESAR project [60] would also be the next step if Garmin watches are to be used.

Bibliography

[1] *Accuracy*. URL: https://www.garmin.com/en-US/legal/atdisclaimer/. (Accessed: 09.03.2022).

[2] Alexandra Amidon. *A Brief Survey of Time Series Classification Algorithms*. URL: https://towardsdatascience.com/a-brief-introduction-to-time-series-classification-algorithms-7b4284d31b97. (Accessed: 22.02.2022).

[3] Gabriel Martins de Barros et al. 'Smartwatch, oxygen saturation, and COVID-19: Trustworthy?' In: *ABCS Health Sciences* 131 (2021), e021101. DOI: 10.7322/abcshs.2020228.1681.

[4] Renesh Bedre. *ANOVA using Python (with examples)*. URL: https://www.reneshbedre.com/blog/anova.html. (Accessed: 10.04.2022).

[5] Richard B. Berry et al. 'Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events'. In: *Journal of clinical sleep medicine* 8.5 (2012), pp. 597–619. DOI: 10.5664/jcsm.2172.

[6] Philip E. Bickler, John R. Feiner and John W. Severinghaus. 'Effects of Skin Pigmentation on Pulse Oximeter Accuracy at Low Saturation'. In: *Anesthesiology* 102 (2005), pp. 715–719. DOI: 10.1097/00000542-200504000-00004.

[7] J. Martin Bland and Douglas G. Altman. 'Statistical methods for assessing agreement between two methods of clinical measurement'. In: *The Lancet* 1 (1986), pp. 307–310. DOI: 10.1016/S0140-6736(86)90837-8.

[8] Jason Brownlee. *What is a Confusion Matrix in Machine Learning*. URL: https://machinelearningmastery.com/confusion-matrix-machine-learning/#:~:text=A%5C%20confusion%5C%20matrix%5C%20is%5C%20a%5C%20summary%5C%20of%5C%20prediction%5C%20results%5C%20on,key%5C%20to%5C%20the%5C%20confusion%5C%20matrix.. (Accessed: 28.04.2022).

[9] Kela Casey. *What is Wearable Technology, How it Works?* URL: https://codersera.com/blog/what-is-wearable-technology-how-it-works/. (Accessed: 25.04.2021).

[10] *CESAR: Using Complex Event Processing for Low-threshold and Non-intrusive Sleep Apnea Monitoring at Home*. URL: https://www.mn.uio.no/ifi/english/research/projects/cesar/index.html. (Accessed: 22.02.2022).

[11] T. Chai and R. R. Draxler. 'Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature'. In: *Geoscientific Model Development* 7 (2014), pp. 1247–1250. DOI: 10.5194/gmd-7-1247-2014.

[12] Chiara Cirelli. *Insufficient sleep: Definition, epidemiology, and adverse outcomes*. URL: https://www.uptodate.com/contents/insufficient-sleep-definition-epidemiology-and-adverse-outcomes. (Accessed: 17.04.2022).

[13] Chiara Cirelli and Giulio Tononi. 'Is Sleep Essential?' In: *PLoS Biology* 6.8 (2008), e216. DOI: 10.1371/journal.pbio.0060216.

[14] Jacob Cohen. 'A Coefficient of Agreement for Nominal Scales'. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[15] *Correlation Coefficient | Types, Formulas & Examples*. URL: https://www.scribbr.com/statistics/correlation-coefficient/. (Accessed: 14.04.2022).

[16] Nurettin Özgür Doğan. 'Bland-Altman analysis: A paradigm to understand correlation and agreement'. In: *Turkish Journal of Emergency Medicine* 18 (2018), pp. 139–141. DOI: 10.1016/j.tjem.2018.09.001.

[17] Jessilynn Dunn, Ryan Runge and Michael Snyder. 'Wearables and the medical revolution'. In: *Personalized Medicine* 15.5 (2018), pp. 429–448. DOI: 10.2217/pme-2018-004.

[18] *Fitzpatrick Scale*. URL: https://emergetulsa.com/wp-content/uploads/2021/10/fitzpatrick-scale-1-800x268-1.png. (Accessed: 15.05.2022).

[19] Kenneth Aune Frisvold. 'Non-Invasive Benchmarking of Pulse Oximeters- An Empirical Approach. Procedures, Considerations and Limitations of Testing Health Sensor Platforms'. MA thesis. Reprosentralen: University of Oslo, 2018.

[20] *Garmin Venu 2S*. URL: https://www.garmin.com/en-US/p/707572#specs. (Accessed: 25.04.2022).

[21] *Garmin vivoactive 4*. URL: https://buy.garmin.com/nb-NO/NO/p/643382#specs. (Accessed: 25.04.2021).

[22] Davide Giavarina. 'Understanding Bland Altman Analysis'. In: *Biochemia Medica* 25.2 (2015), pp. 141–151. DOI: 10.11613/BM.2015.015.

[23] Daniel J. Gottlieb and Naresh M. Punjabi. 'Diagnosis and Management of Obstructive Sleep Apnea A Review'. In: *JAMA* 323 (2020), pp. 1389–1400. DOI: 10.1001/jama.2020.3514.

[24] Dorothy Graham, Rex Black and Erik van Veenendaal. *Foundations of Software Testing: ISTQB Certification*. 4th ed. Cengage Learning EMEA, 2020.

[25] Robert E. Gries and Lee J. Brooks. 'Normal Oxyhemoglobin Saturation During Sleep: How Low Does It Go?' In: *Chest* 110.6 (1996), pp. 1489–1492. DOI: 10.1378/chest.110.6.1489.

[26] Felix Griffin Halvorsen. 'Garmin smartwatches to detect desaturation events as part of OSA screening at home'. MA thesis. Reprosentralen: University of Oslo, 2020.

[27] Ralf E. Harskamp et al. 'Performance of popular pulse oximeters compared with simultaneous arterial oxygen saturation or clinical-grade pulse oximetry: a cross-sectional validation study in intensive care patients'. In: *BMJ Open Respiratory Research* 8 (2021), e000939. DOI: 10.1136/bmjresp-2021-000939.

[28] Marjorie Hecht. *What Are the Fitzpatrick Skin Types?* URL: https://www.healthline.com/health/beauty-skin-care/fitzpatrick-skin-types. (Accessed: 18.04.2022).

[29] Harald Hrubos-Strøm et al. 'A Norwegian population-based study on the risk and prevalence of obstructive sleep apnea'. In: *Journal of Sleep Research* 20 (2010), pp. 162–170. DOI: 10.1111/j.1365-2869.2010.00861.x.

[30] *Medical electrical equipment — Part 2-61: Particular requirements for basic safety and essential performance of pulse oximeter equipment*. Standard. Geneva, CH: International Organization for Standardization, 2017.

[31] *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*. Standard. Geneva, CH: International Organization for Standardization, 2018.

[32] Guðmundur Jónsson. 'Investigating the Application of Semantic Segmentation for Detecting Sleep Apnea in Polygraphy Data'. MA thesis. Reprosentralen: University of Oslo, 2021.

[33] Stein Kristiansen et al. 'A Clinical Evaluation of a Low-Cost Strain Gauge Respiration Belt and Machine Learning to Detect Sleep Apnea'. Unpublished. 2021.

[34] Stein Kristiansen et al. 'Comparing manual and automatic scoring of sleep monitoring data from portable polygraphy'. In: *Journal of Sleep Research* 30 (2020), e13036. DOI: 10.1111/jsr.13036.

[35] Stein Kristiansen et al. 'Data Mining for Patient Friendly Apnea Detection'. In: *IEEE Access* 6 (2018), pp. 74598–74615. DOI: 10.1109/ACCESS.2018.2882270.

[36] Stein Kristiansen et al. 'Machine Learning for Sleep Apnea Detection with Unattended Sleep Monitoring at Home'. In: *Association for Computing Machinery* 2.2 (2021). DOI: 10.1145/3433987.

[37] *Laboratory for Computational Physiology*. URL: https://lcp.mit.edu/physionet. (Accessed: 25.05.2022).

[38] Claire J. Lauterbach et al. 'Accuracy and Reliability of Commercial Wrist-Worn Pulse Oximeter During Normobaric Hypoxia Exposure Under Resting Conditions'. In: *Research Quarterly for Exercise and Sport* (2020), pp. 1–10. DOI: 10.1080/02701367.2020.1759768.

[39] Hooseok Lee, Hoon Ko and Jinseok Lee. 'Reflectance pulse oximetry: Practical issues and limitations'. In: *ICT Express* 2.4 (2016). Special Issue on Emerging Technologies for Medical Diagnostics, pp. 195–198. DOI: 10.1016/j.icte.2016.10.004.

[40] Fredrik Løberg. 'Measuring the Signal Quality of Respiratory Effort Sensors for Sleep Apnea Monitoring. A Metric Based Approach'. MA thesis. Reprosentralen: University of Oslo, 2018.

[41] Fredrik Løberg, Vera Goebel and Thomas Plagemann. 'Quantifying the Signal Quality of Low-cost Respiratory Effort Sensors for Sleep Apnea Monitoring'. In: *3rd International Workshop on Multimedia for Personal Health and Health Care* 3 (2018), pp. 1–9. DOI: 10.1145/3264996.3264998.

[42] Mary L McHugh. 'Interrater reliability: the kappa statistic'. In: *Biochemia Medica* 22.3 (2012), pp. 276–282. DOI: 10.5194/gmd-7-1247-2014.

[43] Fábio Mendonça et al. 'Devices for home detection of obstructive sleep apnea: A review'. In: *Sleep Medicine Reviews* 41 (2018), pp. 149–160. DOI: 10.1016/j.smrv.2018.02.004.

[44] Q. J. W. Milner and G. R. Mathews. 'An assessment of the accuracy of pulse oximeters'. In: *Anaesthesia* 67 (2012), pp. 396–401. DOI: 10.1111/j.1365-2044.2011.07021.x.

[45] Meir Nitzan, Ayal Romem and Robert Koppel. 'Pulse oximetry: fundamentals and technology update'. In: *Medical Devices: Evidence and Research* 7 (2014), pp. 231–239. DOI: 10.2147/MDER.S47319.

[46] *Noxturnal Sleep Study Software – Sleep Software for Sleep Studies*. URL: https://noxmedical.com/products/noxturnal-software/. (Accessed: 09.03.2022).

[47] *OSA blockage*. URL: https://www.clevelandclinic.org/healthinfo/ShowImage.ashx?PIC=4403&width=450. (Accessed: 26.05.2022).

[48] Sai Patwardhan. *Simple understanding and implementation of KNN algorithm!* URL: https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/. (Accessed: 28.04.2022).

[49] Caleb Phillips et al. 'WristO2: Reliable Peripheral Oxygen Saturation Readings from Wrist-Worn Pulse Oximeters'. In: *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. 2021, pp. 623–629. DOI: 10.1109/PerComWorkshops51409.2021.9430986.

[50] *Pulse Oximeter Accuracy and Limitations: FDA Safety Communication*. URL: https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication. (Accessed: 16.04.2022).

[51] Naresh M. Punjabi. 'The Epidemiology of Adult Obstructive Sleep Apnea'. In: *Proceedings of the American Thoracic Society* 5.2 (2008), pp. 136–143. DOI: 10.1513/pats.200709-155MG.

[52] *PureSAT® Advantage*. URL: https://www.nonin.com/wp-content/uploads/2018/09/PureSAT-Advantage-Brochure.pdf. (Accessed: 22.05.2022).

[53] Allison Ragan. *Taking the Confusion Out of Confusion Matrices*. URL: https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e. (Accessed: 28.04.2022).

[54] Nur H. Rashid et al. 'The Value of Oxygen Desaturation Index for Diagnosing Obstructive Sleep Apnea: A Systematic Review'. In: *Laryngoscope* 131 (2021), pp. 440–447. DOI: 10.1002/lary.28663.

[55] Jessica Vensel Rundo and Ralph Downey 3rd. 'Polysomnography'. In: *Journal of Sleep Research* 160 (2019), pp. 162–170. DOI: 10.1016/B978-0-444-64032-1.00025-4.

[56] Anshul Saini. *An Introduction to Random Forest Algorithm for beginners*. URL: https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/. (Accessed: 28.04.2022).

[57] Anshul Saini. *Support Vector Machine(SVM): A Complete guide for beginners*. URL: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/. (Accessed: 28.04.2022).

[58] Ashraf Saleh, Roesnita Binti Isamil and Norasikin Binti Fabil. 'Extension of PACMAD Model For Usability Evaluation Metrics Using Goal Question Metrics (GQM) Approach'. In: *Journal of Theoretical and Applied Information Technology* 79.1 (2015), pp. 90–100.

[59] Jerome M. Siegel. 'Sleep viewed as a state of adaptive inactivity'. In: *Nature Reviews Neuroscience* 10 (2009), pp. 747–753. DOI: 10.1038/nrn2697.

[60] Jagat Deep Singh. 'Nidra: An Extensible Android Application for Recording, Sharing and Analyzing Breathing Data. An Engineering Approach'. MA thesis. Reprosentralen: University of Oslo, 2019.

[61] Michael W. Sjoding et al. 'Racial Bias in Pulse Oximetry Measurement'. In: *The New England Journal of Medicine* 383.25 (2020), pp. 2477–2478. DOI: 10.1056/NEJMc2029240.

[62] *Sleep Diagnostics | Sleep Monitoring Devices | Nox Medical*. URL: https://noxmedical.com/. (Accessed: 25.02.2022).

[63] *Sleep studies*. URL: https://commons.wikimedia.org/wiki/File:Sleep_studies.jpg. (Accessed: 22.05.2022).

[64] Threadcurve Editorial Staff. *14 Top Smartwatch Brands You Need to Know About*. URL: https://threadcurve.com/smartwatch-brands/. (Accessed: 25.04.2021).

[65] *statistics | Hypothesis testing*. URL: https://www.britannica.com/science/statistics/Hypothesis-testing. (Accessed: 14.04.2022).

[66] Eric Suni. *Sleep Apnea*. URL: https://www.sleepfoundation.org/sleep-apnea. (Accessed: 03.11.2019).

[67] Divyanshi Tewari and Asavari Patil. *Smartwatch Market Outlook - 2027*. URL: https://www.alliedmarketresearch.com/smartwatch-market. (Accessed: 25.04.2021).

[68] *The Beginner's Guide to Usability Testing [+ Sample Questions]*. URL: https://blog.hubspot.com/marketing/usability-testing. (Accessed: 13.01.2022).

[69] *The Nox T3s - Next Generation HST Device*. URL: https://noxmedical.com/noxt3s/. (Accessed: 09.03.2022).

[70] Kevin K. Tremper and Steven J. Barker. 'Pulse Oximetry'. In: *Anesthesiology* 70 (1989), pp. 98–108. DOI: 10.1097/00000542-198901000-00019.

[71] *Type I, Type II, Type III Sleep Monitors, CMS AASM Guidelines*. URL: https://clevemed.com/cms-aasm-guidelines-for-sleep-monitors-type-i-type-ii-type-iii/. (Accessed: 26.05.2022).

[72] U.S. Department of Health and Human Services et al. 'Pulse Oximeters - Premarket Notification Submissions [510(k)s]: Guidance for Industry and Food and Drug Administration Staff'. In: *Center for Devices and Radiological Health* (2013).

[73] *Understanding a Pulse Oximeter Report*. URL: https://www.beverlyhillstmjheadachepain.com/sleep-apnea/pulse-oximeter-report/. (Accessed: 30.05.2021).

[74] *Understanding the Results*. URL: https://healthysleep.med.harvard.edu/sleep-apnea/diagnosing-osa/understanding-results. (Accessed: 29.05.2022).

[75] *What is machine learning?* URL: https://www.sap.com/insights/what-is-machine-learning.html. (Accessed: 29.05.2022).

[76] *What Is the Maximum Bluetooth Range of My Garmin Watch or Edge Device?* URL: https://support.garmin.com/en-US/?faq=cRPwF2hIIv0hFYl35tj7T8. (Accessed: 25.04.2021).

[77] Cort J. Willmott and Kenji Matsuura. 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance'. In: *Climate Research* 30 (2005), pp. 79–82. DOI: 10.3354/cr030079.

[78] *WristOx2® Model 3150 with Bluetooth® Low Energy*. URL: https://www.nonin.com/products/wristox2-model-3150-with-ble/. (Accessed: 09.03.2022).

[79] Michael W. Wukitsch et al. 'Pulse oximetry: analysis of theory, technology, and practice'. In: *Journal of clinical monitoring* 4.4 (1988), pp. 290–301. DOI: 10.1007/BF01617328.

Part V

Appendices

Appendix A

Source Code

Following is a link to the repository containing all the work related to this thesis: https://github.uio.no/franceal/francesca-master. A README-file describes the project contents and how to run the different scripts.

Appendix B

Consent Agreement

The consent agreement that had to be signed by the subjects before any data could be collected. There is a Norwegian and English version of the agreement.

CESAR samtykkeerklæring for datainnsamling av fysiologiske data

Evaluering av Garmin smartklokkers evne til å detektere obstruktivt søvnapne gjennom desaturasjon hendelser

Bakgrunn: I CESAR-prosjektet (finansert fra Norges forskningsråd) utfører vi tverfaglig forskning (informatikk og medisin) for å kunne monitorere obstruktivt søvnapne (Obstructive Sleep Apnea - OSA) hjemme for alle pasienter. OSA blir i økende grad anerkjent som en viktig årsak (sykdom) til medisinsk sykelighet og dødelighet. OSA er en relativt vanlig søvnforstyrrelse som er karakterisert ved gjentatte episoder med delvis eller fullstendig kollabering av de øvre luftveier under søvn. Det er estimert at omkring 70-80% av OSA-tilfeller ikke blir diagnostisert. Men også for de som blir diagnostisert tar det ofte for lang tid før en anbefaler en klinisk utredning for OSA fordi symptomene som pasientene beskriver kan ha mange ulike årsaker. Videre er terskelverdien for å starte en OSA-diagnoseprosess veldig høy.

Diagnosen blir vanlighvis gjennomført i et søvnlaboratorium på sykehus ved bruk av polysomnogrrafiinstrumenter med spesielle tester som er ukomfortable for pasienten og som er veldig kostbare (fordi de krever mange ressurser).

Moderne smarttelefoner og billige helsesensorer utgjør en lovende plattform for å samle inn OSA-relatert data hjemme. Målet er å gjøre det enkelt for vanlige brukere å få en pekepinn på om man burde kontakte fastlege for å vurdere om en OSAutredning bør gjennomføres. Formålet er å starte diagnoseprosessen så tidlig som mulig uten å overdiagnostisere pasientene. For å oppnå dette skal vi utvikle ny programvare som bruker maskinlæring og dagens konsumerelektronikk med egnede sensorer for å supplere klassisk polysomnografi. Vi skal undersøke anvendeligheten av veiledet og ikke veiledet maskinlæring for å identifisere viktige/interessante mønster i dataene som kan generere ny kunnskap i OSA-forskningen og for å utvikle verktøy for on-line analyse. Metodene for å designe verktøy for on-line analyse skal være anvendbare også for personer med lite IT-kunnskap slik at de kan gjøre tilpasninger som muliggjør personspesifisk OSA-analyse.

Datainnsamling: Følgende fysiologiske data blir samlet inn for bruk i vår studie: kjønn, hudtype som definert av Fitzpatrick skalaen, oxygenmetning og akselerometer data.

Frivillig deltakelse: All deltagelse er frivillig, og du kan trekke deg når som helst. Du kan når som helst avslutte datainnsamling eller trekke tilbake informasjon som er gitt under observasjon.

Anonymitet: Opptakene (datasett) vil bli anonymisert. Det vil si at ingen

andre enn prosjektmedlemer vil vite hvem som er blitt monitorert, og informasjonen vil ikke kunne tilbakeføres til deg. Før opptakene begynner ber vi deg om å samtykke i deltagelsen ved å undertegne på at du har lest og forstått informasjonen på dette arket og ønsker å delta.

Samtykke: Jeg har lest informasjonen ovenfor og samtykker i at de nevnte opplysningene registreres i CESAR-OSA databasen og gjøres tilgjengelig for forskning i CESAR prosjektet.

Navn, sted, dato og signatur

CESAR agreement for physiological data collection

Evaluating Garmin Smartwatches Ability to Detect Obstructive Sleep Apnea (OSA) Through Desaturation Events

Background: The CESAR project (financed by the Norwegian Research Council) performs interdisciplinary research (computer science and medicine) to enable monitoring of Obstructive Sleep Apnea (OSA) at home for everybody. OSA is being increasingly recognized as an important cause of medical morbidity and mortality. It is a relatively common sleep disorder that is characterized by recurrent episodes of partial or complete collapse of the upper airway during sleep. It is estimated that about 70-80% of OSA cases are not diagnosed. Proper sleep is crucial for maintaining good physical and mental health.

Diagnosing OSA is usually done by hospitalization in sleep laboratories with polysomnographic instruments with multi-parametric tests. The overall process of diagnosing OSA is on the one hand rather uncomfortable for the patients and on the other hand it is very resource demanding in terms of costs of specialized equipment, hospital space, staff for patient support, and expert assessment of polysomnography results. Modern smart phones and low-cost medical/health sensors are a promising platform to collect OSA related data at home. We aim to make it easier for the average user to be able to get an idea if she/he should contact a physician for to perform an OSA examination. To achieve this we will develop new software solutions using machine learning techniques to bridge state-of-the-art consumer electronic devices with appropriate sensors to supplement the classical polysomnography. We will investigate the usefulness of supervised and unsupervised learning (data mining) techniques to identify interesting data patterns that might lead to new knowledge in OSA research and to support the design and engineering of the on-line analysis tool. The design of the on-line analysis tool is driven by the goal to enable individuals with limited computing skill to customize and personalize the on-line analysis.

Collected data: We collect the following physiological data to be used in our research: gender, skin type as defined by the Fitzpatrick scale, oxygen saturation, and accelerometer.

Voluntary participation: All participation is voluntary. You can terminate your participation in the physiological data collection at any time. You can withdraw your information (demand the deletion of your data) at any time.

Anonymity: Your registered data will be anonymized, that means only the project members will know who is monitored, and your information can not be linked to your person. Before the data registration can start, you need to read, understand and sign this agreement that you want to participate in this study.

Agreement: I have read the information above and I agree that the described physiological data will be registered in the CESAR-OSA database and that the information will be used for research in the CESAR project.

Name, place, date and signature

Appendix C

Usability Test Guide

This usability test guide was used for conducting the usability tests of the *Cesar smartwatches* app.

C.1 Welcome and Purpose

Thank you so much for wanting to participate in this usability test. I wanted to give you a little information about what you will be doing and give you time to ask any questions you might have before we get started.

You will serve as an evaluator of an app and complete a set of tasks. Our goal is to see how easy or difficult you find the app to use and if there are any improvements that could be made. I am here to record your experience and comments about the app. During this session, I would like you to think aloud as you work to complete the tasks. I may ask you to clarify what you have said or ask you for information on what you were looking for or what you expect to have happen.

You will have to wear this Garmin Venu 2S smartwatch and the app to be evaluated is called *Cesar Smartwatches* and can be found on this Android phone. You are going to perform the two main tasks that can be done in the app and tell me how easy or difficult they were to perform. There is no right or wrong answers so please use the app as you normally would any other app. If you have any questions, comments or areas of confusion while you are working, please let me know. Try to work through the tasks based on what you see on screen, but if you reach a point where you are not sure what to do then you can ask for assistance.

We will be doing an audio recording of this session for reference if needed. Your name will not be associated or reported with data or findings from this evaluation and you are free to end the session whenever you like. The session will start with some background questions about your phone and app use. I may ask you other questions as we go and we will have wrap up questions at the end.

C.2 Introductory Questions

- Are you familiar with smartwatches? Do you currently own/use a smartwatch? Have you previously used/owned a smartwatch?
- If yes, do you often use apps connected to the watch?
- How would you describe your phone use, especially apps on your phone?

C.3 Tasks

- **First task**: Subject has to pair the Venu 2S watch with the app.

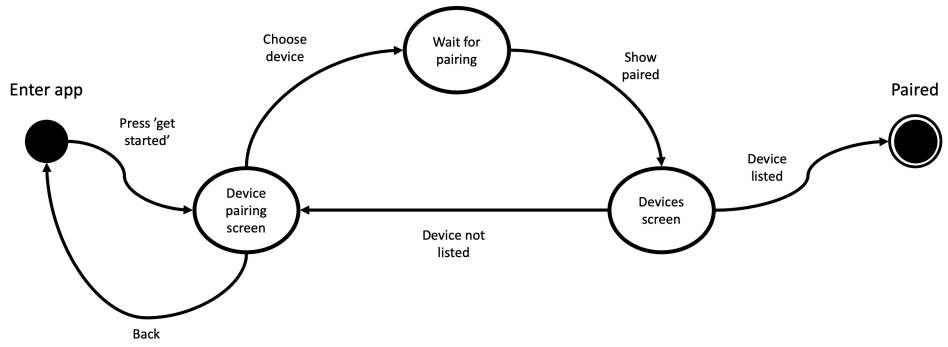


Figure C.1: *State-transition diagram of pairing a device*

  - What was this experience like?
  - Did you find it easy or difficult to pair the watch?
  - Is there something that was confusing/didn't understand?

- **Second task**: Subject has to start and end a new recording.

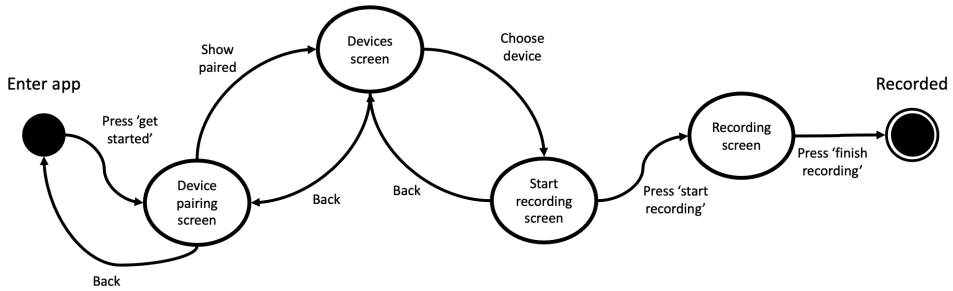


Figure C.2: *State-transition diagram of performing a recording*

  - What was this experience like?
  - Did you find it easy or difficult to record?
  - Is there something that was confusing/didn't understand?

C.4 Exit Questions/Final User Impressions

- What is your overall impression of the app?
- What did you like best about the app?
- What did you like least about the app?
- What would you like to improve with the app if you had the opportunity?
- Is there anything that you feel is missing on the app?
- Do you have any other final comments or questions?
- Are you interested in taking part in a follow-up test with a later version of the app?

Appendix D

Experiment Results

The results of the classification performance metrics for each recording for all classifiers, and all signal quality metrics for each recording. The recordings are sorted from best to worst for the quality metrics table.

D.1 Classification Performance Results

ODI - Normal					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-101	Garmin	-0.119	0.667	0.0	0.909
	Nox	0.0	0.733	0.0	1.0
R-102	Garmin	-0.256	0.5	0.0	0.778
	Nox	0.243	0.714	0.2	1.0
R-103	Garmin	0.025	0.538	0.167	0.857
	Nox	0.0	0.538	0.0	1.0
R-106	Garmin	-0.037	0.429	0.125	0.833
	Nox	0.0	0.429	0.0	1.0
R-109	Garmin	-0.125	0.6	0.0	0.9
	Nox	0.0	0.667	0.0	1.0
R-1013	Garmin	0.34	0.643	0.375	1.0
	Nox	0.0	0.429	0.0	1.0
R-1014	Garmin	0.0	0.786	0.0	1.0
	Nox	0.0	0.786	0.0	1.0
R-1015	Garmin	0.0	0.857	0.0	1.0
	Nox	0.0	0.857	0.0	1.0
Mean	Garmin	-0.021($\pm$0.2)	0.63($\pm$0.1)	0.083($\pm$0.1)	0.91($\pm$0.08)
	Nox	0.03($\pm$0.09)	0.64($\pm$0.2)	0.025($\pm$0.07)	1.0($\pm$0e+00)

Table D.1: *Performance metrics for ODI on normal*

ODI - Tight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-211	Garmin	0.0	0.769	0.0	1.0
	Nox	0.755	0.923	0.667	1.0
R-1115	Garmin	0.44	0.857	0.333	1.0
	Nox	0.0	0.786	0.0	1.0
R-117	Garmin	0.0	0.786	0.0	1.0
	Nox	0.0	0.786	0.0	1.0
R-1114	Garmin	0.0	0.857	0.0	1.0
	Nox	0.0	0.857	0.0	1.0
R-115	Garmin	0.323	0.786	0.25	1.0
	Nox	0.0	0.714	0.0	1.0
R-111	Garmin	0.0	0.571	0.0	1.0
	Nox	0.0	0.571	0.0	1.0
R-112	Garmin	0.0	0.714	0.0	1.0
	Nox	0.0	0.714	0.0	1.0
R-119	Garmin	0.462	0.786	0.4	1.0
	Nox	0.0	0.643	0.0	1.0
Mean	Garmin	0.15($\pm$0.2)	0.77($\pm$0.09)	0.12($\pm$0.2)	1.0($\pm$0e+00)
	Nox	0.094($\pm$0.3)	0.75($\pm$0.1)	0.083($\pm$0.2)	1.0($\pm$0e+00)

Table D.2: *Performance metrics for ODI on tight*

ODI - Back					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1214	Garmin	0.0	0.769	0.0	1.0
	Nox	0.435	0.846	0.333	1.0
R-123	Garmin	0.103	0.643	0.2	0.889
	Nox	0.243	0.714	0.2	1.0
R-122	Garmin	-0.132	0.467	0.0	0.875
	Nox	0.0	0.533	0.0	1.0
R-121	Garmin	0.0	0.643	0.0	1.0
	Nox	0.0	0.643	0.0	1.0
R-221	Garmin	0.0	0.727	0.0	1.0
	Nox	0.0	0.727	0.0	1.0
Mean	Garmin	-0.0058($\pm$0.08)	0.65($\pm$0.1)	0.04($\pm$0.09)	0.95($\pm$0.06)
	Nox	0.14($\pm$0.2)	0.69($\pm$0.1)	0.11($\pm$0.2)	1.0($\pm$0e+00)

Table D.3: *Performance metrics for ODI on back*

ODI - Overnight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1315	Garmin	0.128	0.859	0.412	0.876
	Nox	0.0	0.962	0.0	1.0
R-139	Garmin	0.038	0.912	0.077	0.958
	Nox	0.0	0.948	0.0	1.0
R-1312	Garmin	0.124	0.97	0.333	0.975
	Nox	0.0	0.993	0.0	1.0
R-136	Garmin	-0.083	0.641	0.1	0.828
	Nox	0.0	0.744	0.0	1.0
R-135	Garmin	0.006	0.889	0.018	0.986
	Nox	0.0	0.9	0.0	1.0
R-131	Garmin	0.096	0.836	0.164	0.923
	Nox	0.0	0.886	0.0	1.0
R-132	Garmin	-0.022	0.948	0.0	0.963
	Nox	0.0	0.985	0.0	1.0
R-133	Garmin	0.066	0.917	0.2	0.933
	Nox	0.178	0.98	0.1	1.0
R-236	Garmin	-0.006	0.667	0.042	0.954
	Nox	0.148	0.719	0.117	0.996
R-231	Garmin	-0.027	0.873	0.029	0.948
	Nox	-0.013	0.911	0.0	0.992
R-232	Garmin	-0.014	0.938	0.0	0.945
	Nox	0.0	0.992	0.0	1.0
Mean	Garmin	0.028($\pm$0.07)	0.86($\pm$0.1)	0.12($\pm$0.1)	0.94($\pm$0.05)
	Nox	0.028($\pm$0.07)	0.91($\pm$0.1)	0.02($\pm$0.04)	1.0($\pm$0.003)

Table D.4: *Performance metrics for ODI on overnight*

KNN - Normal					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1015	Garmin	0.0	0.5	0.5	0.5
	Nox	1.0	1.0	1.0	1.0
R-1014	Garmin	0.0	0.5	0.333	0.667
	Nox	0.0	0.5	0.0	1.0
R-1013	Garmin	-0.167	0.417	0.667	0.167
	Nox	0.333	0.667	0.333	1.0
R-102	Garmin	-0.2	0.4	0.6	0.2
	Nox	0.2	0.6	0.2	1.0
R-103	Garmin	0.333	0.667	0.5	0.833
	Nox	0.167	0.583	0.333	0.833
R-101	Garmin	0.0	0.5	0.25	0.75
	Nox	0.5	0.75	0.5	1.0
R-106	Garmin	0.167	0.583	0.667	0.5
	Nox	0.333	0.667	0.667	0.667
R-109	Garmin	-0.4	0.3	0.4	0.2
	Nox	0.0	0.5	0.0	1.0
Mean	Garmin	-0.033($\pm$0.2)	0.48($\pm$0.1)	0.49($\pm$0.2)	0.48($\pm$0.3)
	Nox	0.32($\pm$0.3)	0.66($\pm$0.2)	0.38($\pm$0.3)	0.94($\pm$0.1)

Table D.5: *Performance metrics for KNN on normal*

KNN - Tight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-211	Garmin	0.0	0.5	1.0	0.0
	Nox	0.333	0.667	0.333	1.0
R-1115	Garmin	0.333	0.667	0.667	0.667
	Nox	0.333	0.667	0.333	1.0
R-117	Garmin	0.333	0.667	1.0	0.333
	Nox	0.0	0.5	0.0	1.0
R-1114	Garmin	0.5	0.75	0.5	1.0
	Nox	0.0	0.5	0.5	0.5
R-115	Garmin	0.25	0.625	0.5	0.75
	Nox	0.0	0.5	0.25	0.75
R-111	Garmin	0.167	0.583	0.333	0.833
	Nox	0.5	0.75	0.5	1.0
R-112	Garmin	-0.25	0.375	0.5	0.25
	Nox	0.0	0.5	0.0	1.0
R-119	Garmin	0.2	0.6	0.6	0.6
	Nox	0.2	0.6	0.6	0.6
Mean	Garmin	0.19($\pm$0.2)	0.6($\pm$0.1)	0.64($\pm$0.2)	0.55($\pm$0.3)
	Nox	0.17($\pm$0.2)	0.59($\pm$0.1)	0.31($\pm$0.2)	0.86($\pm$0.2)

Table D.6: *Performance metrics for KNN on tight*

KNN - Back					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1214	Garmin	0.667	0.833	0.667	1.0
	Nox	0.667	0.833	0.667	1.0
R-123	Garmin	0.0	0.5	0.4	0.6
	Nox	0.2	0.6	0.2	1.0
R-122	Garmin	0.714	0.857	0.857	0.857
	Nox	0.143	0.571	0.143	1.0
R-121	Garmin	0.2	0.6	0.6	0.6
	Nox	0.2	0.6	0.4	0.8
R-221	Garmin	0.0	0.5	0.0	1.0
	Nox	0.0	0.5	0.0	1.0
Mean	Garmin	0.32($\pm$0.4)	0.66($\pm$0.2)	0.5($\pm$0.3)	0.81($\pm$0.2)
	Nox	0.24($\pm$0.3)	0.62($\pm$0.1)	0.28($\pm$0.3)	0.96($\pm$0.09)

Table D.7: *Performance metrics for KNN on back*

KNN - Overnight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1315	Garmin	0.118	0.559	0.588	0.529
	Nox	0.588	0.794	0.765	0.824
R-139	Garmin	0.115	0.558	0.577	0.538
	Nox	0.615	0.808	0.654	0.962
R-1312	Garmin	0.0	0.5	0.667	0.333
	Nox	0.333	0.667	0.333	1.0
R-136	Garmin	-0.2	0.4	0.4	0.4
	Nox	0.5	0.75	0.6	0.9
R-135	Garmin	0.109	0.555	0.545	0.564
	Nox	0.4	0.7	0.436	0.964
R-131	Garmin	0.164	0.582	0.545	0.618
	Nox	0.418	0.709	0.436	0.982
R-132	Garmin	-0.4	0.3	0.2	0.4
	Nox	0.2	0.6	0.2	1.0
R-133	Garmin	0.0	0.5	0.7	0.3
	Nox	0.5	0.75	0.5	1.0
R-236	Garmin	0.083	0.542	0.425	0.658
	Nox	0.425	0.712	0.55	0.875
R-231	Garmin	0.235	0.618	0.471	0.765
	Nox	0.471	0.735	0.529	0.941
R-232	Garmin	-0.667	0.167	0.333	0.0
	Nox	0.667	0.833	1.0	0.667
Mean	Garmin	-0.04($\pm$0.3)	0.48($\pm$0.1)	0.5($\pm$0.1)	0.46($\pm$0.2)
	Nox	0.47($\pm$0.1)	0.73($\pm$0.07)	0.55($\pm$0.2)	0.92($\pm$0.1)

Table D.8: *Performance metrics for KNN on overnight*

SVM - Normal					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1015	Garmin	-0.5	0.25	0.5	0.0
	Nox	1.0	1.0	1.0	1.0
R-1014	Garmin	-0.333	0.333	0.333	0.333
	Nox	0.0	0.5	0.333	0.667
R-1013	Garmin	-0.333	0.333	0.5	0.167
	Nox	0.333	0.667	1.0	0.333
R-102	Garmin	0.0	0.5	1.0	0.0
	Nox	0.2	0.6	0.2	1.0
R-103	Garmin	0.333	0.667	0.5	0.833
	Nox	0.667	0.833	0.667	1.0
R-101	Garmin	0.0	0.5	0.0	1.0
	Nox	0.5	0.75	0.5	1.0
R-106	Garmin	0.5	0.75	1.0	0.5
	Nox	0.667	0.833	0.833	0.833
R-109	Garmin	0.2	0.6	0.4	0.8
	Nox	0.2	0.6	0.6	0.6
Mean	Garmin	-0.017($\pm$0.4)	0.49($\pm$0.2)	0.53($\pm$0.3)	0.45($\pm$0.4)
	Nox	0.45($\pm$0.3)	0.72($\pm$0.2)	0.64($\pm$0.3)	0.8($\pm$0.2)

Table D.9: *Performance metrics for SVM on normal*

SVM - Tight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-211	Garmin	0.0	0.5	1.0	0.0
	Nox	0.667	0.833	0.667	1.0
R-1115	Garmin	1.0	1.0	1.0	1.0
	Nox	0.333	0.667	1.0	0.333
R-117	Garmin	0.667	0.833	1.0	0.667
	Nox	0.0	0.5	0.0	1.0
R-1114	Garmin	0.0	0.5	0.5	0.5
	Nox	0.0	0.5	0.5	0.5
R-115	Garmin	-0.5	0.25	0.25	0.25
	Nox	0.75	0.875	1.0	0.75
R-111	Garmin	0.0	0.5	0.167	0.833
	Nox	0.167	0.583	1.0	0.167
R-112	Garmin	0.0	0.5	0.5	0.5
	Nox	0.25	0.625	0.25	1.0
R-119	Garmin	-0.2	0.4	0.6	0.2
	Nox	0.0	0.5	1.0	0.0
Mean	Garmin	0.12($\pm$0.5)	0.56($\pm$0.2)	0.63($\pm$0.3)	0.49($\pm$0.3)
	Nox	0.27($\pm$0.3)	0.64($\pm$0.1)	0.68($\pm$0.4)	0.59($\pm$0.4)

Table D.10: *Performance metrics for SVM on tight*

SVM - Back					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1214	Garmin	0.0	0.5	0.333	0.667
	Nox	0.0	0.5	0.667	0.333
R-123	Garmin	0.4	0.7	1.0	0.4
	Nox	0.6	0.8	0.6	1.0
R-122	Garmin	-0.571	0.214	0.0	0.429
	Nox	0.143	0.571	0.143	1.0
R-121	Garmin	0.0	0.5	1.0	0.0
	Nox	-0.2	0.4	0.8	0.0
R-221	Garmin	0.0	0.5	0.0	1.0
	Nox	0.333	0.667	0.333	1.0
Mean	Garmin	-0.034($\pm$0.3)	0.48($\pm$0.2)	0.47($\pm$0.5)	0.5($\pm$0.4)
	Nox	0.18($\pm$0.3)	0.59($\pm$0.2)	0.51($\pm$0.3)	0.67($\pm$0.5)

Table D.11: *Performance metrics for SVM on back*

SVM - Overnight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1315	Garmin	0.059	0.529	0.588	0.471
	Nox	0.647	0.824	0.882	0.765
R-139	Garmin	0.115	0.558	0.654	0.462
	Nox	0.577	0.788	0.654	0.923
R-1312	Garmin	0.333	0.667	0.667	0.667
	Nox	0.667	0.833	0.667	1.0
R-136	Garmin	0.2	0.6	0.6	0.6
	Nox	0.6	0.8	0.8	0.8
R-135	Garmin	0.0	0.5	0.073	0.927
	Nox	0.491	0.745	0.618	0.873
R-131	Garmin	0.145	0.573	0.491	0.655
	Nox	0.382	0.691	0.455	0.927
R-132	Garmin	-0.4	0.3	0.0	0.6
	Nox	0.6	0.8	0.8	0.8
R-133	Garmin	0.0	0.5	0.5	0.5
	Nox	0.4	0.7	1.0	0.4
R-236	Garmin	0.1	0.55	0.583	0.517
	Nox	0.533	0.767	0.875	0.658
R-231	Garmin	0.0	0.5	0.147	0.853
	Nox	0.353	0.676	0.588	0.765
R-232	Garmin	0.0	0.5	0.0	1.0
	Nox	0.667	0.833	1.0	0.667
Mean	Garmin	0.05($\pm$0.2)	0.53($\pm$0.09)	0.39($\pm$0.3)	0.66($\pm$0.2)
	Nox	0.54($\pm$0.1)	0.77($\pm$0.06)	0.76($\pm$0.2)	0.78($\pm$0.2)

Table D.12: *Performance metrics for SVM on overnight*

RF - Normal					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1015	Garmin	0.5	0.75	0.5	1.0
	Nox	0.5	0.75	1.0	0.5
R-1014	Garmin	-0.333	0.333	0.333	0.333
	Nox	0.0	0.5	0.333	0.667
R-1013	Garmin	-0.167	0.417	0.667	0.167
	Nox	0.167	0.583	0.5	0.667
R-102	Garmin	0.0	0.5	0.8	0.2
	Nox	0.4	0.7	0.4	1.0
R-103	Garmin	0.0	0.5	0.167	0.833
	Nox	0.5	0.75	0.667	0.833
R-101	Garmin	-0.5	0.25	0.0	0.5
	Nox	0.25	0.625	0.5	0.75
R-106	Garmin	0.0	0.5	0.5	0.5
	Nox	0.0	0.5	0.667	0.333
R-109	Garmin	0.2	0.6	0.4	0.8
	Nox	0.2	0.6	0.4	0.8
Mean	Garmin	-0.038($\pm$0.3)	0.48($\pm$0.2)	0.42($\pm$0.3)	0.54($\pm$0.3)
	Nox	0.25($\pm$0.2)	0.63($\pm$0.1)	0.56($\pm$0.2)	0.69($\pm$0.2)

Table D.13: *Performance metrics for RF on normal*

RF - Tight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-211	Garmin	0.333	0.667	1.0	0.333
	Nox	0.667	0.833	0.667	1.0
R-1115	Garmin	0.333	0.667	0.667	0.667
	Nox	0.667	0.833	0.667	1.0
R-117	Garmin	0.667	0.833	1.0	0.667
	Nox	0.0	0.5	0.0	1.0
R-1114	Garmin	0.0	0.5	0.5	0.5
	Nox	0.0	0.5	1.0	0.0
R-115	Garmin	0.25	0.625	0.5	0.75
	Nox	0.75	0.875	1.0	0.75
R-111	Garmin	0.167	0.583	0.5	0.667
	Nox	0.5	0.75	0.833	0.667
R-112	Garmin	-0.25	0.375	0.5	0.25
	Nox	0.5	0.75	0.5	1.0
R-119	Garmin	0.0	0.5	0.6	0.4
	Nox	0.6	0.8	1.0	0.6
Mean	Garmin	0.19($\pm$0.3)	0.59($\pm$0.1)	0.66($\pm$0.2)	0.53($\pm$0.2)
	Nox	0.46($\pm$0.3)	0.73($\pm$0.1)	0.71($\pm$0.3)	0.75($\pm$0.3)

Table D.14: *Performance metrics for RF on tight*

RF - Back					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1214	Garmin	0.0	0.5	0.333	0.667
	Nox	0.667	0.833	0.667	1.0
R-123	Garmin	0.0	0.5	0.8	0.2
	Nox	0.6	0.8	0.6	1.0
R-122	Garmin	-0.143	0.429	0.286	0.571
	Nox	0.143	0.571	0.143	1.0
R-121	Garmin	0.2	0.6	0.6	0.6
	Nox	0.2	0.6	0.6	0.6
R-221	Garmin	-0.333	0.333	0.0	0.667
	Nox	0.667	0.833	0.667	1.0
Mean	Garmin	-0.055($\pm$0.2)	0.47($\pm$0.1)	0.4($\pm$0.3)	0.54($\pm$0.2)
	Nox	0.46($\pm$0.3)	0.73($\pm$0.1)	0.54($\pm$0.2)	0.92($\pm$0.2)

Table D.15: *Performance metrics for RF on back*

RF - Overnight					
	Device	κ	**Accuracy**	**Sensitivity**	**Specificity**
R-1315	Garmin	0.118	0.559	0.588	0.529
	Nox	0.706	0.853	0.941	0.765
R-139	Garmin	0.038	0.519	0.654	0.385
	Nox	0.577	0.788	0.692	0.885
R-1312	Garmin	0.0	0.5	0.667	0.333
	Nox	0.333	0.667	0.333	1.0
R-136	Garmin	-0.3	0.35	0.5	0.2
	Nox	0.5	0.75	0.8	0.7
R-135	Garmin	0.218	0.609	0.636	0.582
	Nox	0.564	0.782	0.727	0.836
R-131	Garmin	0.218	0.609	0.636	0.582
	Nox	0.473	0.736	0.636	0.836
R-132	Garmin	0.2	0.6	0.4	0.8
	Nox	0.8	0.9	1.0	0.8
R-133	Garmin	0.5	0.75	0.9	0.6
	Nox	0.5	0.75	0.8	0.7
R-236	Garmin	0.175	0.588	0.575	0.6
	Nox	0.392	0.696	0.808	0.583
R-231	Garmin	0.0	0.5	0.441	0.559
	Nox	0.441	0.721	0.824	0.618
R-232	Garmin	0.0	0.5	0.333	0.667
	Nox	0.333	0.667	0.333	1.0
Mean	Garmin	0.11($\pm$0.2)	0.55($\pm$0.1)	0.58($\pm$0.2)	0.53($\pm$0.2)
	Nox	0.51($\pm$0.1)	0.76($\pm$0.07)	0.72($\pm$0.2)	0.79($\pm$0.1)

Table D.16: *Performance metrics for RF on overnight*

D.2 Signal Quality Results

	Accuracy	MAE	Mean Bias	Precision	Upper LoA	Lower LoA
R-1010*	0.975	0.733	0.353	1.817	2.134	-1.427
R-104	1.305	1.133	-0.720	2.178	1.414	-2.854
R-117	1.315	1.025	-0.966	1.784	0.782	-2.715
R-1015	1.519	1.340	-0.794	2.589	1.744	-3.332
R-201	1.647	1.304	-0.117	3.285	3.103	-3.336
R-127	1.682	1.356	-0.416	3.260	2.779	-3.611
R-108	1.701	1.349	0.276	3.356	3.565	-3.014
R-1215	1.849	1.624	-1.376	2.471	1.046	-3.797
R-114	1.854	1.587	-1.369	2.500	1.081	-3.819
R-1011	1.889	1.566	-1.490	2.324	0.787	-3.767
R-211	2.008	1.141	0.069	4.013	4.002	-3.863
R-136	2.017	1.560	0.728	3.762	4.415	-2.959
R-1115	2.094	1.676	-0.252	4.158	3.823	-4.327
R-122	2.120	1.686	1.204	3.489	4.624	-2.215
R-135	2.159	1.487	0.337	4.265	4.517	-3.842
R-109	2.279	1.685	-0.551	4.423	3.784	-4.885
R-115	2.291	1.956	1.280	3.799	5.003	-2.442
R-101	2.348	2.101	2.091	2.137	4.185	-0.004
R-113	2.349	1.974	-1.157	4.087	2.848	-5.163
R-1214	2.357	2.073	-1.611	3.441	1.761	-4.983
R-102	2.383	1.560	0.661	4.578	5.148	-3.825
R-1013	2.465	2.016	0.024	4.931	4.856	-4.808
R-221	2.468	2.190	1.178	4.337	5.428	-3.073
R-139	2.679	2.209	-1.370	4.605	3.142	-5.883
R-132	2.769	2.278	1.683	4.396	5.992	-2.625
R-119	2.788	2.176	-0.704	5.395	4.583	-5.992
R-121	2.827	2.294	-2.183	3.594	1.339	-5.705
R-107	2.852	2.516	2.306	3.357	5.596	-0.984
R-133	2.870	2.251	0.809	5.507	6.206	-4.589
R-111	2.937	2.333	-0.661	5.724	4.948	-6.270
R-1014	3.065	2.413	-2.097	4.472	2.286	-6.479
R-236	3.146	2.528	-0.588	6.181	5.469	-6.645
R-103	3.155	2.616	2.039	4.815	6.758	-2.680
R-1312	3.286	2.632	1.667	5.664	7.218	-3.883
R-1114	3.441	3.298	-3.298	1.963	-1.373	-5.222

R-124	3.851	2.703	0.959	7.459	8.269	-6.351
R-231	3.983	2.487	2.255	6.566	8.689	-4.180
R-1012	4.079	3.339	3.284	4.839	8.026	-1.459
R-106	4.552	3.637	1.972	8.206	10.014	-6.071
R-1315	4.823	3.670	3.189	7.237	10.281	-3.904
R-131	5.135	3.952	3.850	6.796	10.510	-2.810
R-232	5.416	4.506	3.671	7.964	11.476	-4.134
R-105**	5.609	5.515	5.515	2.038	7.512	3.518
R-123	7.582	6.206	6.015	9.231	15.062	-3.031
R-112	9.883	7.359	6.806	14.332	20.851	-7.240
Mean	3.01($\pm$1.6)	2.39($\pm$1.3)	0.62($\pm$2.1)	4.73($\pm$2.3)	5.26($\pm$4.1)	-4.02($\pm$1.6)

Table D.17: *Signal quality metrics for all experiments. Light gray → normal experiment, medium → tight experiment, dark → back of the wrist and white → overnight monitoring. *Venu 2S recorded only 150 SpO_2 signals, not included. **No variation in oximetry data detected by Venu 2S, not included*