KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Multimodal Signal Analysis for Unobtrusive Characterization of Obstructive Sleep Apnea

Margot Deviaene

Supervisors:
Prof. dr. ir. S. Van Huffel
Prof. dr. D. Testelmans
Prof. dr. ir. C. Varon

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

October 2020

Multimodal Signal Analysis for Unobtrusive Characterization of Obstructive Sleep Apnea

Margot DEVIAENE

Examination committee:
Prof. dr. ir. Y. Willems, chair
Prof. dr. ir. S. Van Huffel, supervisor
Prof. dr. D. Testelmans, supervisor
Prof. dr. ir. C. Varon, supervisor
Prof. dr. ir. A. Bertrand
Prof. dr. B. Buyse
Prof. dr. ir. J. Bergmans
(TU Eindhoven)
Prof. dr. T. Penzel
(Charité University Hospital Berlin)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

October 2020

© 2020 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Margot Deviaene, Kasteelpark Arenberg 10 box 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

Four years and a half ago, I was in the middle of my master thesis and thinking what to do next, when my promotor Prof. Sabine Van Huffel presented me the opportunity to start a PhD in her group. I was hesitant at first, because I had not planned to stick around in Leuven, but in the end I was persuaded by the challenging and interesting research topic. Sabine, I would like to thank you for having confidence in me and to give me the opportunity to start my PhD in BIOMED. Moreover, thank you for all the support throughout the years, providing feedback and advice which steered my research in the right direction. I would also like to thank you for creating the wonderful atmosphere in the BIOMED group by organizing numerous social events. The BIOMED reflection day made me feel welcome in the group from the start. You also welcomed us together with our partners into your home for the pizza evening and Christmas party which is much appreciated.

I would also like to thank my co-promotor Prof. Carolina Varon for her guidance throughout my PhD. Thank you Carolina for encouraging me to keep going at difficult times through your enthusiasm, and by seeing research opportunities when it felt to me that I hit a dead end. Even when you left for Delft, you still kept in touch and were available whenever I needed you.

The collaboration with the sleep lab also had a major impact on my research. I would like to thank my co-promotor Prof. Dries Testelmans, Prof. Bertien Buyse and Pascal Borzée for the nice collaboration and to inspire me by showing how my work could be useful in clinical practice. Thank you for all the effort put in the collection of the data sets. A special thanks goes to Pascal and Johan for guiding the wearable experiments.

I would also like to thank the other members of the examination committee Chairman Prof. Yves Willems, Prof. Alexander Bertrand, Prof. Jan Bergmans and Prof. Thomas Penzel for their interest in my work and for accepting the invitation to be part of this jury. Thank you for your feedback and the

interesting discussions during my preliminary defense.

Most of the research within my PhD was performed within the OSA+ project. I would like to thank all collaborators within this project for their input on our progress meetings. Especially thanks to Jan, Gabrielle Bruno and Pedro for the nice research discussions, and to Sebastiaan, Merel, Hans and Roy for the clinical input from Kempenhaeghe and to provide us the data set of healthy volunteers. Also thanks to Jesús and Dorien, my BIOMED colleagues involved in this project which have helped me improve my research. I also had the pleasure to collaborate with visiting researcher Javier Milagro from the University of Zaragoza, and Ivan Castro of Imec. Thank you for these nice collaborations.

The last collaborators I would like to thank are all my colleagues from the BIOMED group. Thank you for all your input and research discussions. But most of all, thank you for the nice and supporting work atmosphere in which we celebrated all of each others successes and each opportunity was good enough to bring cakes or cookies. The last few months during the COVID-19 teleworking period, I have missed our interesting lunch conversations, the small chitchats and sharing our struggles and successes. Also outside the office we bonded a lot during conferences and social activities organized within the group. There were a lot of fun parties, which started off for me with a very memorable one being the birthday party of Griet and Rob. I also have fond memories of my sleep themed bachelorette party, the successes of our (girls') football team at the ESAT sport days, our crypto-coin adventure which started off very promising, lasertag, Sunday funday, and so many more fun activities organized within the group. I would want to thank every one of you for creating these nice memories. Special thanks to my office mates Laure, Griet, Jesús, Dorien and Nick for the nice office atmosphere.

Tenslotte zou ik ook nog mijn familie, schoonfamilie en vrienden willen bedanken voor alle steun de afgelopen jaren, voor de bemoedigende woorden en interesse in mijn werk. En natuurlijk ook voor de welgekomen afleiding en ontspanning door middel van uistapjes, feestjes, weekendjes en etentjes. Mama en papa, dankjewel om mij van jongs af aan te steunen in alles wat ik doe, en me de vrijheid te geven om mijn eigen pad te kiezen. Om altijd voor me klaar te staan als ik jullie hulp nodig heb en om een warme thuis te voorzien waarnaar ik altijd terug op bezoek kan komen. Kevin, bedankt om mijn proefschrift na te lezen, en voor al je liefde en steun. Ook al was ik tijdens de laatste weken van mijn doctoraat niet echt de aangenaamste persoon om mee samen te leven, toch bleef je voor mij klaar staan. Dankjewel om mij de moeilijke perioden te helpen overwinnen en mij de moed te geven om mijn doctoraat af te ronden.

Margot

Abstract

Obstructive sleep apnea (OSA) is the most prevalent sleep related breathing disorder, nevertheless subjects suffering from it often remain undiagnosed due to the cumbersome diagnosis procedure. Moreover, the prevalence of OSA is increasing and a better phenotyping of patients is needed in order to prioritize treatment. The goal of this thesis was to tackle those challenges in OSA diagnosis. Additionally, two main algorithmic contributions which are generally applicable were proposed within this thesis. The binary interval coded scoring algorithm was extended to multilevel problems and novel monotonicity constraints were introduced. Moreover, improvements to the random-forest based feature selection were proposed including the use of the Cohen's kappa value, patient independent validation, and further feature pruning steered by the correlation between features.

These novel methods were applied together with classification and feature selection methods from the literature to improve the OSA diagnosis. All methods were predominantly tested on data sets collected at the sleep laboratory of the UZ Leuven. Within this PhD, I contributed to the collection and structuring of these data sets to have them ready-to-use for collaborators and future generations of students.

The first application was the development of reliable, multimodal OSA screening methods based on unobtrusive measurements such as oxygen saturation (SpO_2), electrocardiography (ECG), pulse photoplethysmography (PPG), and respiratory measures. An SpO_2-based model was proposed which extracted features from oxygen desaturations, including novel estimates of quasi-periodicity. These were then used to identify desaturations caused by apneas using a random forest classifier. Furthermore, the use of a robust linear regression was introduced to estimate the apnea-hypopnea index (AHI) from the number of detected apneic desaturations per hour of recording. The obtained model was tested on several independent test sets and obtained OSA screening accuracies larger than 88 % over all sets.

Unimodal methods based on the ECG and PPG were proposed as well. The ECG-based model included extra cardiorespiratory measures and investigated the use of the slope range as an alternative ECG-derived respiration (EDR). In addition, the method was designed such that the ECG and EDR signals could easily be replaced by unobtrusive ECG and respiratory measures. The PPG model was based on the extraction of time series known to be modulated by the respiration and/or the autonomous nervous system. Additionally, the use of instantaneous estimates of the heart rate variability using a point-process model was investigated.

Different multimodal OSA detection approaches were explored. The inclusion of the respiratory effort signal in the ECG-based algorithm led to an improvement in performance. While the straightforward addition of the PPG and ECG features to the SpO_2-based model was not effective. The use of a 4-state hidden Markov model was investigated as a smarter way to combine the SpO_2 and PPG features. Additionally, the impact of sleep staging on the performance of these models was tested.

Finally, a main contribution of this PhD was to test the developed ECG and PPG OSA detection algorithms on unobtrusive signals, including capacitively-coupled ECG and bioimpedance, and wearable PPG recordings. Both experiments showed promising results, especially using both capacitively-coupled signals, but also limitations of the current algorithms on the unobtrusive data.

In the last part of this thesis, a contribution towards a better characterization of OSA patients beyond the AHI was proposed. Novel pulse oximetry markers were developed and investigated to assess the cardiovascular status of OSA subjects. It was found that patients with cardiovascular comorbidities experienced more severe oxygen desaturations and incomplete resaturations to the baseline SpO_2 values. These markers were shown to be of added value to patient demographics to assess the cardiovascular status of OSA patients. The proposed PPG features, on the other hand, did not improve the patient classification.

The obtained SpO_2 markers are a major contribution of this PhD as they might give more physiological insights in the detrimental effects of apneas on the cardiovascular system, and they might be useful to prioritize OSA patients for treatment. These parameters were further investigated by performing statistical tests to compare the different cardiovascular conditions and the influence of drug use. Moreover, the novel multilevel interval coded scoring was used to train a model to predict the cardiovascular status of OSA patients. The final model included the age, BMI and the SpO_2 parameters, and obtained good classification performances on a clinical population. In order to further validate the predictive power of these novel SpO_2 markers, a clinical OSA data set including a long-term cardiovascular follow-up should be used.

Beknopte samenvatting

Obstructieve slaapapneu (OSA) is de meest voorkomende slaapgerelateerde ademhalingsstoornis. Nochtans wordt de ziekte vaak niet gediagnosticeerd en blijven veel patiënten onbehandeld. Bovendien neemt de prevalentie van OSA toe en is een betere fenotypering van patiënten nodig om prioriteiten te stellen in de behandeling. Het doel van deze thesis was om deze uitdagingen in de diagnose van OSA aan te pakken. Daarnaast zijn twee belangrijke algoritmische bijdragen geïntroduceerd die algemeen toepasbaar zijn. Het binaire interval gecodeerde scoremodel werd uitgebreid naar multi niveau problemen en extra monotoonheidsbeperkingen werden ingevoerd. Bovendien werden verbeteringen aan de random forest parameter selectie voorgesteld, waaronder het gebruik van de Cohen's kappa-waarde, patiëntonafhankelijke validatie en de verdere reductie op basis van de correlatie tussen parameters.

Deze nieuwe methoden werden samen met classificatie- en parameter selectiemethoden uit de literatuur toegepast om de OSA diagnose te verbeteren. De methoden werden voornamelijk getest op datasets opgenomen in het slaaplaboratorium van het UZ Leuven. Binnen dit doctoraat heb ik bijgedragen tot het verzamelen en ordenen van deze datasets om ze gebruiksklaar te maken.

De eerste toepassing was de ontwikkeling van betrouwbare, multimodale OSA-screeningmethoden op basis van signalen die opgenomen kunnen worden op een comfortabele manier, zoals de zuurstofsaturatie (SpO_2), elektrocardiogram (EKG), fotoplethysmografie (PPG) en ademhalingssignalen. Een SpO_2-gebaseerd model werd geïntroduceerd dat parameters berekend op basis van de zuurstofdesaturaties, inclusief nieuwe schattingen van quasi-periodiciteit. Deze werden dan gebruikt om de desaturaties veroorzaakt door apneu's te identificeren met behulp van een random forest model. Bovendien werd het gebruik van robuuste lineaire regressie voorgesteld om de apneu-hypopneu index (AHI) te schatten op basis van het aantal apneu gerelateerde desaturaties per uur. Het verkregen model werd getest op verschillende onafhankelijke testsets en behaalde een OSA-screening accuraatheid groter dan 88 %.

Unimodale methoden op basis van het EKG en de PPG werden ook voorgesteld. Het EKG-model onderzocht extra cardiorespiratoire parameters en het gebruik van de slope range als alternatieve EKG-afgeleide ademhaling (EDR). Bovendien werd de methode zo ontworpen dat de EKG en EDR-signalen gemakkelijk vervangen konden worden door contactloze EKG en ademhalingsmetingen. Het PPG model was gebaseerd op tijdsreeksen die beïnvloed worden door de ademhaling en/of het autonome zenuwstelsel. Daarnaast werd het gebruik van ogenblikkelijke schattingen van de hartslagvariabiliteit met behulp van een point-process model onderzocht.

Vervolgens werden verschillende multimodale OSA-detectie methoden uitgetest. De toevoeging van het ademhalingsinspanningssignaal aan het EKG model leidde tot een verbeterde performantie. De toevoeging van de PPG en EKG parameters aan het SpO_2 model, was niet effectief. Het gebruik van een hidden Markov model met 4 toestanden werd onderzocht als slimmere combinatie van de SpO_2 en PPG parameters. Tot slot werd de impact van slaapstadia op de performantie van deze modellen onderzocht

Een belangrijke bijdrage van deze thesis was het uittesten van de ontwikkelde EKG en PPG OSA-detectiealgoritmes op meer comfortabele opnametechnieken zoals capacitief gekoppelde EKG en bio-impedantiesignalen, en draagbare PPG opnames. Beide experimenten resulteerden in veelbelovende resultaten, vooral voor het gebruik van beide capacitieve signalen, maar brachten ook beperkingen van de huidige algoritmes op deze signalen naar boven.

In het laatste deel van deze thesis werd een bijdrage voorgesteld voor een betere karakterisering van OSA-patiënten. Nieuwe pulse oximetrie parameters werden ontwikkeld en onderzocht voor de beoordeling van de cardiovasculaire status. Patiënten met cardiovasculaire comorbiditeiten bleken ernstigere zuurstofdesaturaties en onvolledige resaturaties naar de basis SpO_2 waarden te ondervinden. Deze parameters hadden een toegevoegde waarde bij het beoordelen van de cardiovasculaire status van OSA-patiënten. De voorgestelde PPG parameters daarentegen, verbeterden de classificatie niet. De verkregen SpO_2 parameters zijn een belangrijke bijdrage van dit doctoraat aangezien ze fysiologische inzichten kunnen verschaffen omtrend de schadelijke effecten van apneu's, en nuttig kunnen zijn om prioriteiten te stellen in de behandeling. Deze parameters werden verder onderzocht door statistische testen uit te voeren die de cardiovasculaire aandoeningen vergeleken en de invloed van medicatiegebruik testten. Bovendien werd het nieuwe multi-niveau interval gecodeerd scoringsmodel toegepast om de cardiovasculaire status te voorspellen. Het finale model was gebaseerd op leeftijd, BMI en SpO_2 parameters, en behaalde een goede performantie op een klinische populatie. Om de voorspellende kracht van deze nieuwe SpO_2 parameters te valideren, moet een klinische dataset met een cardiovasculaire opvolging op de lange termijn gebruikt worden.

List of Abbreviations

AASM	American Academy of Sleep Medicine.
AC	Alternating current.
Acc	Accuracy.
ACF	Autocorrelation function.
ACM	Accelerometry.
AF	Atrial fibrillation.
AHI	Apnea-hypopnea index.
AUC	Area under the reiceiver operating characteristic curve.
BCG	Ballistocardiography.
BMI	Body mass index.
cC	Cardiovascular conditions.
ccBioz	Capacitively-coupled bioimpedance.
ccECG	Capacitively-coupled electrocardiogram.
cE	Cardiovascular events.
CNN	Convolutional neural network.
CPAP	Continuous positive airway pressure.
CR	Cardiorespiratory.
Ct	Cardiovascular controls.
CV	Cardiovascular.
DAP	Decreases in the amplitude fluctuation of the pulse photoplethysmography.
DC	Direct current.
ECG	Electrocardiography.
EDR	Electrocardiogram derived respiration.
EEG	Electroencephalography.

EMG	Electromyogram.
EOG	Electrooculogram.
ESS	Epworth sleepiness scale.
FN	False negative.
FP	False positive.
FPR	False positive rate.
FS LS-SVM	Fixed-size least-squares support vector machine.
GMM	Gaussian mixture model.
GUI	Graphical user interface.
Hb	Hemoglobin.
HF	High frequency range, 0.15-0.4 Hz.
HMM	Hidden Markov model.
HR	Heart rate.
HRV	Heart rate variability.
ICS	Interval coded scoring.
IR	Infrared.
kNN	k-nearest neighbors.
LDA	Linear discriminant analysis.
LED	Light emitting diode.
LF	Low frequency range, 0.04-0.15 Hz.
LOPO CV	Leave-one-patient-out cross validation.
LS-SVM	Least-squares support vector machine.
MA	Moving average.
MAD	Mandibular advancement devices.
mICS	Multilevel interval coded scoring.
mRMR	Minimum redundancy maximum relevance.
NPV	Negative predictive value.
NSRR	National sleep research resource.
O_2Hb	Oxyhemoglobin.
ODI	Oxygen desaturation index.
OOB	Out-of-bag.
OSA	Obstructive sleep apnea.

PALM	Pcrit, Arousal threshold, Loop gain and Muscle responsiveness.
PAT	Peripheral arterial tonometry.
PI	Predictor importance.
PLM	Periodic limb movements.
PNS	Parasympathetic nervous system.
PPG	Pulse photoplethysmography.
PPV	Positive predictive value.
PR	Pulse rate.
PRSA	Phase rectified signal averaging.
PSD	Power spectral density.
PSG	Polysomnography.
Ramp	R-wave amplitude.
RBF	Radial basis function.
REM	Rapid-eye movement.
RF	Random forest.
RIP	Respiratory inductance plethysmography.
RMSSD	Root mean square of interbeat differentials.
RNN	Recurrent neural network.
ROC	Receiver operating characteristic.
RSA	Respiratory sinus arrythmia.
SCOPER	Sleep, Cardiovascular, Oximetry, Position, Effort and Respiratory.
SDSD	Standard deviation of interbeat intervals.
Se	Sensitivity.
SHHS	Sleep heart health study.
SNS	Sympathetic nervous system.
Sp	Specificity.
SpO_2	Oxygen saturation.
SR	Slope range.
std	Standard deviation.
SVM	Support vector machines.
T90	Time spent below an oxygenation level of 90 %.
TB	Thoracic respiratory effort belt.
TN	True negative.
TP	True positive.

List of Symbols

κ	Cohen's kappa performance metric
$\mathbf{x}_i$	Input data vector for the ith sample
μ	Mean
π_k	Prior class probability of class k
σ	Standard deviation
ξ, e, ε	Slack values
c, γ	Regularization parameters
$f_{k(x)}$	Class density function of class k, depending on the input space x
$I(\mathbf{x}; \mathbf{y})$	Mutual information between variable $\mathbf{x}$ and $\mathbf{y}$
R^2	Coefficient of determination
$\boldsymbol{\alpha}$	Vector with the Lagrange multipliers of an optimization problem
$\mathbf{w}$	Weights of a linear combination
$\varphi(x)$	Feature mapping applied to x
ξ_i	SVM slack variables
b	Bias of a linear combination
$K(\mathbf{x}, \mathbf{z})$	Kernel function evaluated for the vectors $\mathbf{x}$ and $\mathbf{z}$
y_i	Class label of the ith sample

Contents

Abstract iii

Beknopte samenvatting v

List of Abbreviations ix

List of Symbols xi

Contents xiii

List of Figures xix

List of Tables xxiii

I Background 1

1 Introduction 2
1.1 Research motivation 2
1.2 Obstructive sleep apnea 3
1.2.1 Definition of OSA 3
1.2.2 Cardiovascular comorbidities 5
1.2.3 Diagnosis 6
1.2.4 Limitations of current diagnosis 8
1.3 Research objectives 9
1.4 Chapter overview 9
1.5 Collaborations 14

2 Data collection and labeling 17
2.1 Description of the studied signal modalities 17
2.1.1 Oximetry: SpO_2 and PPG 17

2.1.2 Electrocardiography signals . . . 20
2.1.3 Respiratory signals . . . 22
2.2 Cardiovascular labels . . . 23
2.3 Overview of the used data sets . . . 23
2.3.1 Data sets recorded at UZ Leuven . . . 24
2.3.2 Sleep heart health study . . . 27
2.3.3 Physionet Apnea-ECG . . . 27
2.3.4 Kempenhaeghe controls . . . 28

3 Literature review: Detection, screening and phenotyping of OSA 29
3.1 Introduction on OSA screening . . . 30
3.2 Cardiovascular . . . 32
3.2.1 ECG . . . 32
3.2.2 PPG . . . 34
3.2.3 PAT . . . 34
3.2.4 Contactless cardiovascular measurements . . . 34
3.3 Oximetry . . . 35
3.4 Effort and Respiration . . . 37
3.4.1 Airflow . . . 37
3.4.2 Respiratory effort . . . 37
3.4.3 Contactless measures of respiration . . . 38
3.5 Sound . . . 38
3.6 Automatically generated features using deep learning . . . 39
3.7 Multimodal OSA detection . . . 39
3.8 Comparison of methods and modalities . . . 40
3.9 Overview of OSA screening devices . . . 41
3.10 Future of sleep apnea screening: beyond the AHI? . . . 42

II Algorithms 45

4 Classification 46
4.1 Introduction on classification methods . . . 46
4.2 Feature-based classifiers . . . 47
4.2.1 k-nearest neighbors . . . 47
4.2.2 Linear discriminant analysis . . . 48
4.2.3 Random forest . . . 48
4.2.4 (Least-squares) Support vector machines . . . 49
4.2.5 Interval coded scoring . . . 51
4.3 Multilevel ICS . . . 53
4.3.1 Extension to multilevel ICS . . . 53
4.3.2 mICS model with monotonicity constraints . . . 54
4.4 Comparison of methods . . . 55

4.5 Performance evaluation . . . 56
4.5.1 Confusion matrix . . . 56
4.5.2 Receiver operating characteristic curve . . . 57
4.5.3 kappa score . . . 57
4.5.4 Training, validation and test set . . . 58
4.6 Conclusion . . . 59

5 Feature selection **61**
5.1 Introduction to feature selection . . . 61
5.2 Feature selection methods from the literature . . . 62
5.2.1 F-score . . . 62
5.2.2 Minimal redundancy maximal relevance criterion . . . 63
5.2.3 Wrapper methods . . . 63
5.3 Random forest based feature selection . . . 64
5.3.1 Patient independent validation . . . 65
5.3.2 Cohen's kappa value as performance estimate . . . 65
5.3.3 Inclusion of feature correlation . . . 65
5.3.4 Determination of the optimal number of features . . . 66
5.4 Qualitative comparison of feature selection methods . . . 66
5.5 Conclusion . . . 67

III Apnea Detection **69**

6 SpO_2-based sleep apnea detection **70**
6.1 Data sets . . . 71
6.2 Methods . . . 71
6.2.1 SpO_2 preprocessing . . . 72
6.2.2 Desaturation event and baseline extraction . . . 72
6.2.3 Class label computation for the desaturation events . . . 74
6.2.4 Feature extraction . . . 75
6.2.5 Feature selection and classification . . . 77
6.2.6 Statistical analysis of classification performance . . . 78
6.2.7 AHI estimation . . . 79
6.3 Results . . . 79
6.3.1 Feature selection . . . 79
6.3.2 Classification performance . . . 81
6.3.3 Estimation of the AHI . . . 82
6.3.4 OSA classification . . . 84
6.4 Discussion . . . 84
6.4.1 Apnea event detection . . . 84
6.4.2 AHI computation and OSA screening . . . 88
6.5 Additional experiments . . . 91

6.5.1 Application of RF-based feature selector 91
6.5.2 Retraining classifiers on UZ Leuven data sets 95
6.5.3 OSA screening using a mICS model 98
6.6 Conclusion 100

7 ECG-based sleep apnea detection 103
7.1 Materials 104
7.2 Methods 104
7.2.1 Data preprocessing 104
7.2.2 Training apnea detection classifier 107
7.3 Results 109
7.3.1 Extracted features 109
7.3.2 Apnea detection 109
7.4 Discussion 110
7.4.1 Selected features 110
7.4.2 Apnea detection 111
7.5 Conclusion 112

8 PPG-based sleep apnea detection 113
8.1 Materials 113
8.2 Methods 114
8.2.1 PPG pre-processing and feature extraction 114
8.2.2 Feature Selection 115
8.2.3 Classification 115
8.3 Results 116
8.4 Discussion 116
8.5 Conclusion 118

9 Multimodal sleep apnea detection 119
9.1 OSA detection based on ECG and respiratory effort 120
9.1.1 Methods 120
9.1.2 Results 120
9.2 Addition of PPG and ECG features to SpO_2-based apnea detection 121
9.2.1 Materials and methods 122
9.2.2 Results 123
9.3 Apnea detection using HMM based on SpO_2 and PPG features 124
9.3.1 Hidden Markov model for sleep apnea detection 125
9.3.2 Materials and methods 126
9.3.3 Results 127
9.4 Added value of sleep staging for OSA screening 128
9.4.1 Methods 129
9.4.2 Results 129
9.5 Conclusion 133

10 Wearable sleep apnea detection **135**
10.1 Capacitively-coupled ECG and respiration for the detection of sleep apnea . . . 135
10.1.1 Methods . . . 136
10.1.2 Results . . . 140
10.1.3 Discussion . . . 143
10.2 Apnea detection using a reflective Byteflies PPG sensor . . . 145
10.2.1 Methods . . . 145
10.2.2 Results . . . 147
10.3 Conclusion . . . 151

IV Cardiovascular Comorbidities in Sleep Apnea **153**

11 Pulse oximetry markers for cardiovascular disease in sleep apnea **154**
11.1 Introduction to OSA phenotyping . . . 154
11.2 Materials . . . 155
11.3 Methods . . . 155
11.3.1 Extraction of pulse oximetry markers . . . 155
11.3.2 Feature selection and classification . . . 157
11.4 Results . . . 157
11.5 Discussion . . . 157
11.6 Conclusion . . . 161

12 Multilevel ICS to assess the cardiovascular status of sleep apnea patients using oxygen saturation markers **163**
12.1 Materials . . . 164
12.2 Methods . . . 164
12.2.1 Oxygen saturation parameters . . . 164
12.2.2 Multi-level ICS for the assessment of CV comorbidities . . . 167
12.2.3 Statistical analysis . . . 168
12.3 Results . . . 168
12.3.1 Extracted SpO_2 parameters . . . 168
12.3.2 mICS Model . . . 170
12.3.3 Differences between cardiovascular disorders . . . 174
12.3.4 Influence of drug use . . . 174
12.4 Discussion . . . 175
12.4.1 SpO_2 parameters and cardiovascular comorbidities . . . 175
12.4.2 mICS model for cardiovascular comorbidity assessment . . . 176
12.4.3 Predictive power of the obtained mICS model . . . 177
12.4.4 Influence of training set selection on the performance . . . 179
12.4.5 Limitations of the study . . . 179
12.5 Conclusion . . . 180

13 Conclusions and future directions **181**
13.1 Conclusions 181
13.1.1 Multimodal OSA screening methods 182
13.1.2 Characterization of OSA patients beyond the AHI . . . 184
13.1.3 Algorithmic contributions 185
13.2 Future directions 185
13.2.1 Multimodal OSA screening methods 185
13.2.2 Characterization of OSA patients beyond the AHI . . . 187

A SpO_2 based apnea detection GUI **189**

Bibliography **193**

Curriculum vitae **207**

List of publications **209**

List of Figures

1.1 Overview of the cardiovascular consequences of apneic events [114]. 6
1.2 Chapter-by-chapter overview of the structure of the manuscript. 11

2.1 Working principles of an oximeter. Figure adapted from [124]. . 18
2.2 Light absorbance spectra for O_2Hb and Hb. 19
2.3 Byteflies sensor dot. 20
2.4 Characteristic waveform of a normal ECG [129]. 21
2.5 Recording setup of the mattress with non-contact sensors embedded, measuring ccECG and ccBioz. 22

3.1 Example of a segment with apneas. 31
3.2 Desaturation characteristics which can be extracted from oxygen desaturations. 35
3.3 Periodicity in the SpO_2 signal. 36

4.1 Example of finding the optimal hyperplanes for a multilevel problem, data extracted from the Fisher's Iris data set. 54
4.2 Receiver operating characteristic curve [43]. 57

5.1 Schematic overview of the random forest-based feature selection. 65

6.1 Overview of the extraction of desaturations in the SpO_2 signal. 73
6.2 Extraction of desaturation severity features. 74
6.3 Overview of the different steps in the feature selection methodology. 77
6.4 Features selected for SpO_2-based apnea detection. 80
6.5 Scatter and Bland-Altman plot of the SpO_2-based AHI estimate. 83
6.6 Confusion matrix for SpO_2-based OSA screening. 85
6.7 Apnea detection performance per patient in function of the AHI. 87
6.8 Overview of the classifier performance values, according to the number of features that were removed for RF-based feature selector. 92
6.9 3D plot of the selected oximetry features for one patient. 94

6.10 Overview of AHI estimation on UZ Leuven OSA+ test set with retrained SpO_2 model. 98

7.1 Overview of the applied ECG signal preprocessing methods. . . 105
7.2 Development of the ECG apnea detection classifier. 107
7.3 Selected features using the Ramp EDR on the Imec data set. . 109

8.1 Overview of the PPG peak point delineation and feature extraction. 114
8.2 Selected PPG features for normal breathing (0) and apneic segments (1) on the OSA+ data set. 116

9.1 Respiratory and CR interaction features using the TB on the Imec data set. 121
9.2 Overview of the HMM states and their transition probabilities. 128
9.3 Estimated vs true AHI value for SpO_2-based OSA detection with and without including sleep staging on the Imec data set. . . . 130
9.4 Estimated vs true AHI value for PPG-based OSA detection with and without including sleep staging on the Imec data set. 131
9.5 Estimated vs true AHI value for ECG SR EDR based OSA detection with and without including sleep staging on the Imec data set. 131
9.6 Estimated vs true AHI value for ECG TB based OSA detection with and without including sleep staging on the Imec data set. 132

10.1 Overview of the applied signal preprocessing methods. 137
10.2 Example of preprocessed ccECG and ccBioz signals on a 1-minute segment including an apneic event. 138
10.3 Overview of the used signal modality combinations for testing. 139
10.4 Overview of the additional experiments. 139
10.5 ECG features extracted from the Imec set using the PSG ECG and ccECG. 140
10.6 Log kurtosis of the respiration extracted from the imec set using the different respiratory measurements. 141
10.7 Log WAV_{HF} feature extracted from the Imec set using different signal modality combinations. 141
10.8 Overview of the steps in the synchronization of the wearable PPG signals to the PSG. 146
10.9 10 minute PPG signal sample from the PSG and Byteflies recordings, with the extracted PR on the left. 148
10.1015 second PPG signal sample from the PSG and byteflies recordings. 149
10.11Extracted features from the IR channel wearable PPG signals. 149

11.1 Overview of the selected features to assess the CV status. . . . 158

11.2 Posterior probability of the SVM classifier. 160

12.1 Overview of the applied methods. 165
12.2 Comparison of SpO_2 parameters between a cardiovascular control and event patient with similar AHI. 166
12.3 Overview of the extracted SpO_2 parameters for the different cardiovascular comorbidity groups. 169
12.4 Resulting mICS model including SpO_2 parameters, age and BMI. 172
12.5 mICS model risks for the different data sets and cardiovascular group. 173
12.6 Confusion matrix of classifying Leuven test subjects using the mICS model including SpO_2 parameters, age and BMI. 173

A.1 Graphical user interface to estimate the AHI and mICS cardiac risk score parameters from the SpO_2. 191
A.2 Example of SpO_2 plotting in the GUI. 192

List of Tables

2.1 Overview of the used data sets. 25

3.1 Overview of commonly used ECG time-domain features. 33

4.1 Confusion matrix for a two-class problem. 56

6.1 Overview of the extracted SpO_2 features. 75
6.2 Overview of the averaged classification performance on the test data sets. 81
6.3 Overview of the averaged classification performance of the random forest classifier on each data set. 82
6.4 Number of detected desaturations per hour and percentage apneic. 88
6.5 Comparison of OSA patient classification performance for AHI >15 on SHHS2. 89
6.6 Desaturation classification performance results when using the RF-based feature selection method. 92
6.7 Overview of the apnea screening performance obtained using the RF-based feature selection. 93
6.8 Overview of the training data used to retrain models on the UZ Leuven data sets 95
6.9 Overview of the desaturation classification performance for the retrained model on UZ Leuven OSA+ test. 96
6.10 Overview of OSA patient classification performance of retrained models for AHI >15 on UZ Leuven OSA+. 97

7.1 Overview of the number of hours of clean ECG data. 106
7.2 Apneic segment classification performance based on a combination of ECG and EDR estimates. 110
7.3 Performance comparison to the study by Varon et al. [135]. . . 110

8.1 Classification performance of the PPG apnea detection algorithm. 117

9.1 Apneic segment classification performance on the retrained classifier based on the combination of ECG and TB signals. . . . 121
9.2 Classification performance of the multimodal apnea detection models. . . . 124
9.3 HMM state estimation performance. . . . 128
9.4 R^2 values for the linear regression between the estimated and true AHI values for the different signal modalities. . . . 130
9.5 OSA screening performances with an AHI threshold of 15 for the different signal modalities with and without removing wake segments for the OSA+ test set. . . . 132

10.1 Overview of the number of hours of clean data for each signal modality of the Imec data set. . . . 137
10.2 Apneic segment classification performance on the Imec data set for segments with clean capacitive sensor data. . . . 142
10.3 Classification performance of the PPG apnea detection algorithm on the Byteflies PPG data of the Imec data set. . . . 150

11.1 Patient demographics per data set and for each CV comorbidity class. . . . 156
11.2 Cardiovascular comorbidity detection results between the Ct and cE classes. . . . 159

12.1 Patient demographics per data set for the cardiovascular control (Ct), condition (cC) and event (cE) groups. . . . 165
12.2 Classification performance of the different mICS models and Framingham score for the detection of cardiovascular comorbidities. 171

Part I

Background

Chapter 1

Introduction

1.1 Research motivation

Sleep is an important aspect of life, as humans spend on average one third of their life sleeping. Although the function of sleep is not yet completely understood, the negative effects of a lack of sleep are clear. Short term sleep deprivation can cause mood changes, an increased appetite, memory loss and a decrease in productivity and concentration. Whereas, chronic sleep deprivation has been linked to weight gain, diabetes, cardiovascular (CV) diseases, depression, an impaired immune system, and an increased all cause mortality [77]. Nevertheless, many people do not get the recommended daily amount of sleep. This is often caused by societal pressure or lifestyle choices. But also many people suffer from sleep disorders, which disturb their normal sleep.

Obstructive sleep apnea (OSA) is the most common sleep related breathing disorder affecting almost 1 billion people worldwide [14]. It presents itself as both partial and complete cessations of breathing during sleep. These events often result in a drop in blood oxygenation and/or an arousal, causing the normal breathing to be restored. Despite the high prevalence of OSA, the disease frequently remains undiagnosed and consequently untreated. This is partly due to the expensive and cumbersome way of diagnosing, which is usually performed using an overnight in-hospital procedure where several physiological parameters are measured, called polysomnography (PSG). Based on the PSG, sleep stages and respiratory events are afterwards manually scored by a sleep expert according to a set of scoring rules. The PSG is a single night snapshot of the patient's status, which cannot be used for long-term follow-up, nor can it be

measured at home. Therefore, it is not suited for screening in larger populations. Alternative screening and diagnostic tools based on signals that can easily be acquired in a home environment should thus be developed in order to follow-up known OSA patients and to reach a wider population for OSA screening.

The severity of OSA is nowadays measured using the apnea-hypopnea index (AHI), which counts the number of scored respiratory events per hour of sleep. OSA is diagnosed when a patient has an AHI of 5 or higher in combination with symptoms (such as daytime sleepiness or snoring), or an AHI larger than 15, regardless of the presence of symptoms [115]. Although the AHI is a very simple and commonly used diagnostic index, studies have reported several drawbacks. Such as the high prevalence of an AHI larger than 15 in certain demographic populations, the fact that the AHI is independent of the severity of the respiratory events, and the lack of correlation between AHI and cardiovascular risk in OSA patients [56, 60, 70]. Therefore, researchers have been investigating how to characterize OSA beyond the AHI in order to improve the understanding of the underlying pathophysiology, and to help prioritize patients for treatment, reducing their morbidity and mortality.

The goal of this research is to tackle these challenges in OSA diagnosis, first to develop a reliable, multimodal OSA screening method, and second, to define markers for the characterization of OSA patients beyond the AHI. The next section will present the necessary background information on OSA, followed by a more detailed discussion of these research goals.

1.2 Obstructive sleep apnea

1.2.1 Definition of OSA

Pathophysiology

Sleep apnea causes complete cessations or partial reductions of breathing during sleep, called respectively, apneas and hypopneas. These events can be due to a blockage in the upper airway (obstructive) or due to a loss of respiratory drive from the brains (central), although obstructive apneas are more common [1]. The anatomy of the upper airway is an important factor contributing to the development of OSA. Subjects with a narrower upper airway will have a higher predisposition of having OSA. But, also the upper airway dilator muscle activity and responsiveness plays a role [45]. During wake, these muscles will keep the airway open, but during sleep the muscle activity will be diminished. In combination with an unfavorable upper airway anatomy, this can lead to

an upper airway collapse and apneic events. This explains why patients only experience apneas during sleep and why apneas are more prone to occur during rapid-eye movement (REM) sleep, when the muscle tone is at its lowest [45].

Apneic events regularly result in a drop in blood oxygenation and/or an arousal. Although these arousals are assumed to enable patients to breathe again, they also cause sleep fragmentation, leading to daytime sleepiness. Moreover, it has been shown that some OSA patients can restore their breathing without an arousal, or even that for these patients arousals occur after the re-opening of the airway. They suffer from a low arousal threshold and awaken easily from mild apneic events [46].

Another factor that promotes the occurrence of apneic events is an unstable ventilatory control mechanism. This has been described using the concept of loop gain [45]. Some OSA patients have been shown to have an increased respiratory loop gain, meaning that their respiratory system is highly sensitive to breathing disturbances, leading to an excessive response which will not result in normal breathing to be restored. Instead, cyclic breathing patterns might develop in which there is a succession of apneic events and arousals.

In [45], a reduced lung volume is given as a last physiological factor posing an increased risk to develop OSA. Furthermore, the sleeping position also has an influence on the occurrence of apneas. In patients with mild to moderate OSA, having less than 15 events per hour of sleep, it has been shown that apneas often occur predominantly in the supine position [81].

Epidemiology

OSA is two to three times more prevalent in men than in women, and in people older than 65 years [19]. Additionally, obesity is a key risk factor for OSA as fat deposition in the neck increases the collapsibility of the upper airway. The upper airway muscles will also be negatively impacted and fat deposition around the abdomen will lead to a decreased lung volume, increasing the chances of developing OSA [45]. Fortunately, modest weight loss has shown to reduce OSA severity. However, OSA also occurs in people with a normal weight, for whom other factors might contribute to its development, such as the upper airway anatomy, smoking habits, or genetic factors [19].

Treatment

The standard treatment of moderate to severe OSA, having more than 15 apneic events per hour, consists of continuous positive airway pressure (CPAP) therapy.

By administering a constant pressure to the upper airway, a constant air flow to the lungs is imposed and airway collapse is avoided [106].

For subjects with a less severe form of OSA, alternative treatment options are available. Including weight loss and lifestyle changes, such as limiting the intake of alcohol and sedatives [106]. When the OSA is caused mainly due to the anatomy of the upper airway, surgical techniques might be considered. For milder forms, oral appliances might suffice, of which the mandibular advancement devices (MAD) are the most commonly used [106]. In case of positional OSA, positional therapy is also proposed in order to avoid the supine sleeping position [81].

Moreover, recent studies have investigated alternative treatment methods which could be personalized per subject, based on the phenotypical traits causing OSA. These include, for example, sedatives for patients suffering from a low arousal threshold [87].

Symptoms

The most common symptoms of OSA are snoring, due to upper airway obstruction and anomalies, and daytime sleepiness caused by the recurrent arousals distorting the sleep. In addition to those symptoms, patients with OSA have an increased risk of developing cardiovascular comorbidities [19, 83, 141], which will be discussed in the next section.

1.2.2 Cardiovascular comorbidities

Patients with OSA have been shown to have an increased risk of developing hypertension, diabetes and cardiovascular events, such as myocardial infarction and stroke, even when adjusted for confounding factors such as age and obesity [19]. The oxygen desaturations, arousals, and negative intrathoracic pressure swings which accompany apneic events have been linked to a long-term deterioration of the cardiovascular system via different pathways, see Figure 1.1. The exact association is not yet fully understood, but changes in both parasympathetic and sympathetic activity have been recorded, as well as a decreased cardiac output, inflammation and endothelial dysfunction [19, 114].

An observational study by Marin et al. with a mean follow-up period of 10 years, showed that patients with untreated severe OSA have an adjusted odds ratio of 3.17 for developing non-fatal cardiovascular events with respect to healthy subjects [83]. The patient group treated with CPAP, however, had no significant increase in risk. These results suggest that treatment with CPAP can decrease

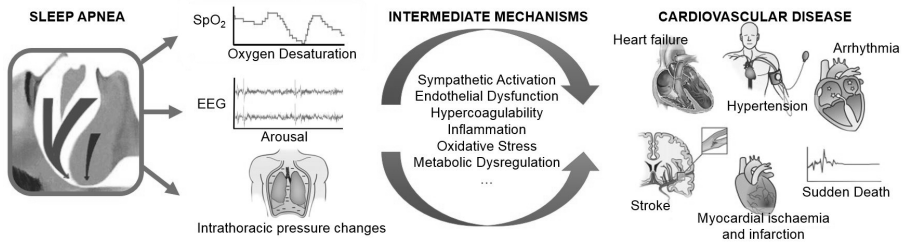


Figure 1.1: Overview of the cardiovascular consequences of apneic events. Apneas cause oxygen desaturations, arousals and intrathoracic pressure changes, which will trigger intermediate mechanisms, leading to several cardiovascular diseases in the long term. Figure adapted from [114].

the risk of developing cardiovascular events. Therefore, it is important to identify high risk patients in order to avert severe cardiovascular comorbidities and mortality by timely OSA diagnosis and treatment.

1.2.3 Diagnosis

Polysomnography

The gold standard for diagnosing OSA is a full night PSG in an attended hospital setting. Multiple physiological signals are recorded as part of the PSG in order to obtain a full characterization of the person's sleep. Below, an overview of the different signals usually measured during a PSG is given.

- **Electroencephalography (EEG):** An EEG with a reduced electrode set recording frontal, central and occipital derivations is recommended during PSG [136]. These EEG measurements are used for sleep stage scoring based on the observed patterns and wave characteristics. Additionally, the EEG will be used to score arousals.
- **Electrooculogram (EOG):** The EOG measures eye-movements, which are useful for the detection of sleep onset and rapid eye movements characteristic for REM sleep.
- **Electromyogram (EMG):** A chin EMG will be used to measure muscle tone, which aids in the sleep stage classification. Moreover, EMG electrodes will also be placed on the limbs in order to assess periodic limb movements (PLM).

- **Respiration:** Airflow measurements are the primary signals used to annotate apneic events in a PSG. The gold standard measurement of airflow is done using a pneumotachograph, which is too uncomfortable to be used during sleep. Therefore, oronasal thermistors and nasal pressure sensors are used to measure airflow during sleep [17].
- **Respiratory effort:** The respiratory effort is measured using respiratory inductance plethysmography (RIP) belts placed around the chest and the thorax. The absence of respiratory effort during central events compared to obstructive events will help to differentiate between the different apnea subtypes.
- **Oximetry:** A finger pulse oximeter is used to measure the oxygen saturation (SpO_2) signal which is used to detect drops in oxygen saturation in the blood due to apneic events. The oximeter will also provide a pulse photoplethysmography (PPG) measurement.
- **Electrocardiography (ECG):** A modified lead-II ECG measurement is generally recorded during a PSG to assess the patient's cardiac function [136].
- **Position:** A body position sensor is included to define if the subject is sleeping in lateral, supine or prone position.
- **Audio:** An audio recording is often performed during the PSG to measure snoring and to verify the nature of arousals.
- **Video:** Video recordings are used to monitor the behavior of the subject and are especially useful for subjects with parasomnias. They can also help to analyze the origin of artefacts in the signals.

Apnea scoring rules

Sleep stages and respiratory events are, in most countries, manually scored using these PSG recordings. In order to ensure uniformity among scoring of sleep studies, the American Academy of Sleep Medicine (AASM) developed a set of scoring rules [1, 17, 64]. According to the AASM 2012 rules, all respiratory events should last longer than 10 seconds, and when the airflow amplitude decreases with more than 90 % the event is scored as an apnea. Hypopnea events, on the other hand, only need a decrease of at least 30 %, accompanied by either an arousal or an oxygen desaturation of at least 3 %.

Some of the older data sets available for this PhD thesis were scored according to the Chicago criteria [107]. According to these rules, respiratory events longer

than 10 seconds are scored when either an airflow drop larger than 50 % occurs, or when a clear airflow amplitude reduction is observed accompanied by an arousal or an oxygen desaturation of at least 3 %.

For the remainder of this thesis, the term apneas will be used as a global term representing both apneas and hypopneas, unless a clear distinction is made between the two.

OSA severity

The severity of OSA is generally measured using the AHI, representing the number of annotated events divided by the hours asleep. As such, OSA severity is divided in four categories: normal (AHI<5), mild (5≤AHI<15), moderate (15≤AHI<30) and severe OSA (AHI≥30) [1].

1.2.4 Limitations of current diagnosis

OSA diagnosis based on an in-hospital PSG has many drawbacks, the procedure is time-consuming, uncomfortable and expensive, which results in long waiting times. This further discourages people to get tested. Moreover, the PSG is not suited for screening of large populations, nor can it be performed multiple nights in a row. Additionally, the manual scoring of events introduces inter and intra scorer variability. The development of automated OSA screening techniques is, therefore, an important field of research as sleep studies need to be brought out of the hospital environment into the subject's home. As such, a long-term follow-up of patients can be performed, and a larger population can be reached in order to find undiagnosed OSA patients and to treat them before they develop OSA related comorbidities. Many researchers have focused on developing alternative screening and diagnostic tools based on signals that can easily be acquired in a home environment. A literature review on this topic can be found in Chapter 3.

Not only the PSG itself has drawbacks, but also the role of the AHI as an OSA severity measure has been questioned. This is because the AHI only counts the number of events, and does not take into account the type of event, nor the severity of the event. The event severity could be defined in terms of duration, degree of oxygen desaturation or presence of an arousal. Another limitation of the AHI is the lack of discriminative ability, which was pointed out by the Hypnolaus study [60]. This study observed that in the general population, up to 50 % of middle-aged men had an AHI larger than 15. Screening this population solely based on the AHI would thus lead to over-treatment, since many a-

symptomatic subjects with an AHI larger than 15 might not benefit from CPAP treatment. The AHI also does not correlate well with OSA symptoms, such as sleepiness measured using the Epworth sleepiness scale (ESS)[121], nor does it correlate well with the cardiovascular risk in OSA patients [56, 70]. Therefore, there is a need to phenotype OSA patients beyond the AHI. Many different directions have been proposed to tackle this problem [143]. Studies have, for example, investigated whether clinical clusters could be defined based on clinical presentation, risk factors and comorbidities [8], or whether a differentiation could be made based on physiological traits causing OSA [44]. In this work, a methodology to differentiate OSA patients based on their cardiovascular status is proposed.

1.3 Research objectives

The goal of this research was to tackle the aforementioned challenges in OSA diagnosis. The first main objective was to develop a reliable, multimodal OSA screening method based on unobtrusive measurements. To this end, several signal modalities were considered, such as SpO_2, ECG, PPG and respiratory measurements. The algorithms were developed and trained based on PSG data sets and afterwards tested on recordings from unobtrusive devices, including Byteflies PPG patches, capacitively-coupled ECG (ccECG) and capacitively-coupled bioimpedance (ccBioz) signals.

The second objective was to define markers for the characterization of OSA patients beyond the AHI. For this, the presence of cardiovascular comorbidities was considered as the main outcome parameter to differentiate patients. The use of markers extracted from the SpO_2 and PPG signals was considered.

1.4 Chapter overview

This thesis is divided into four parts. In the first part, the necessary background information is provided in order to better understand the following chapters. Chapter 2 describes the data sets used for this research, which will be followed by a literature overview on existing OSA detection methods in Chapter 3. The second part of this thesis will focus on the different algorithms which were developed and used in this work. Chapter 4 and 5 will give an overview of, respectively, the classification and feature selection methods. In the third part, these methods will be applied to develop algorithms for the automatic detection of OSA. This part is divided into five chapters, focusing first on

uni-modal approaches using the SpO_2, ECG and PPG signals (see Chapters 6-8), followed by Chapter 9 on multimodal approaches. In Chapter 10, the use of these methods on wearable measurements will be discussed. The final part of this thesis will look beyond the AHI to investigate oximetry markers for the assessment of cardiovascular comorbidities in OSA. This part is divided in two chapters. Chapter 11 focuses on the extraction and selection of SpO_2 and PPG markers. While in Chapter 12, the selected SpO_2 markers will be further analyzed and a multilevel ICS model will be developed to asses the cardiovascular risk. Finally, the manuscript will be concluded with a final chapter providing overall conclusions and future research opportunities.

A graphical overview of the structure of the thesis is given in Figure 1.2 and a more detailed chapter-by-chapter description of the manuscript can be found below.

Part I: Background

Chapter 1

The current chapter introduced the research motivation, background information on OSA including the challenges within the field, and the research objectives. It will be concluded with an overview of the research collaborations within the PhD thesis.

Chapter 2

This chapter will give an overview of the different data sets that were collected in the course of this PhD and that will be used in this research. The collection and structuring of these data sets is a major contribution of this PhD that will be useful for collaborators and future generations of students. This chapter will also describe the acquisition principles of the different signal modalities. Additionally, for some data sets, the presence of cardiovascular comorbidities was assessed at the time of the PSG. This chapter will include a description of the protocol used for this CV assessment.

Chapter 3

In this chapter, a literature review is presented on the detection, screening and phenotyping of sleep apnea in an ambulatory setting.

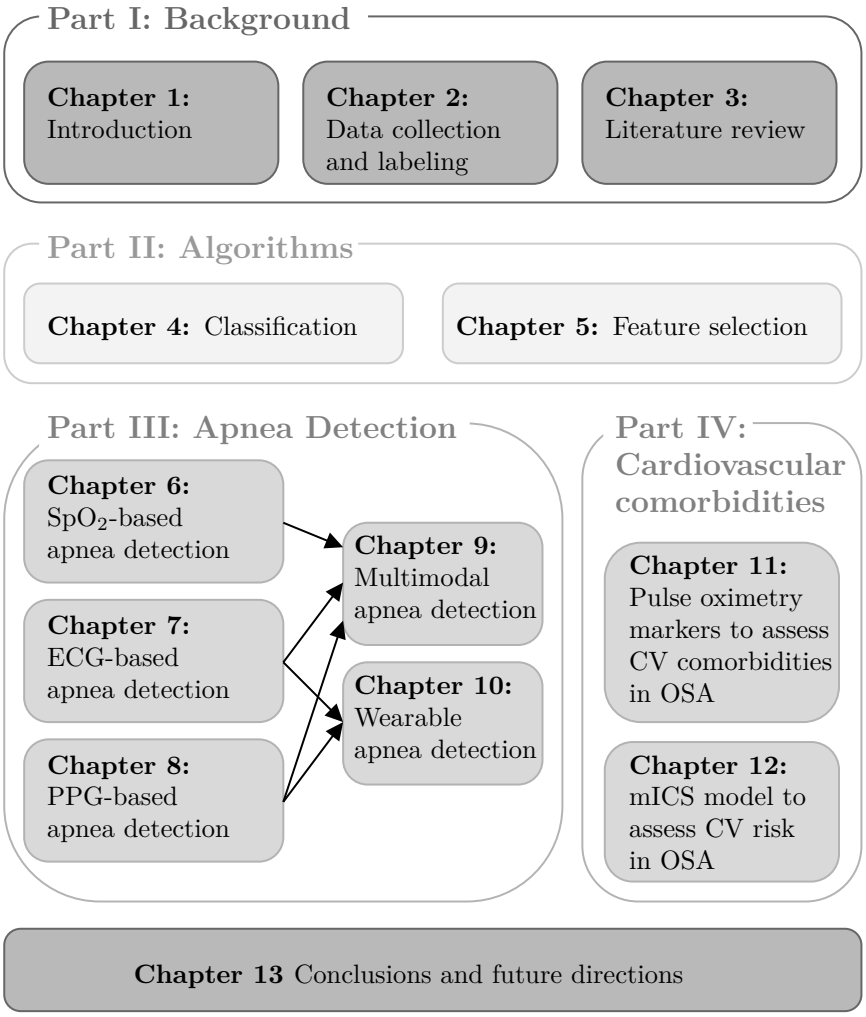


Figure 1.2: Chapter-by-chapter overview of the structure of the manuscript.

Part II: Algorithms

Chapter 4

Classification methods from the literature, including k-nearest neighbors (kNN), linear discriminant analysis (LDA), random forest (RF), support vector machines (SVM) and interval coded scoring (ICS) will be introduced in this chapter.

Moreover, an extension to multilevel ICS (mICS) will be presented, which is one of the main algorithmic contributions of this research. Additionally, the use of extra monotonicity constraints was proposed. These extensions to the ICS toolbox are generally applicable in classification problems where the interpretability of the model is important, therefore, they were published online and are freely available [49].

The chapter will continue with a comparison between the different classification methods and finally, the performance evaluation of classifiers will be discussed.

Chapter 5

This chapter will give an overview of feature selection methods used in this work: the F-score, minimum redundancy maximum relevance (mRMR) and wrapper methods. Followed by the presentation of another contribution of this research, namely, a novel random forest based feature selection method, which was presented at the IEEE EMBS 2019 conference [40]. At the end of the chapter, the different feature selection methods will be compared.

Part III: Apnea Detection

Chapter 6

In this chapter, an SpO_2-based apnea detection algorithm developed during this PhD will be presented, which was published in IEEE JBHI [41]. The method detects all oxygen desaturations, for which then features are extracted that are used to build a classifier to decide if a desaturation is linked to a respiratory event or not. Based on the number of detected desaturations classified as linked to respiratory events, an AHI estimate is computed.

Additionally, extra experiments that were performed after the publication of this study will be presented. These include the application of the RF-based feature selector introduced in Chapter 5, the retraining of the algorithm based on clinical data from the UZ Leuven data sets, and the use of the mICS model for OSA severity classification.

Chapter 7

Chapter seven proposes an algorithm for the automatic detection of sleep apnea based on single lead ECG. It builds further on the methods presented by Varon et

al. in [135], but includes an additional ECG derived respiration (EDR) and novel estimates of cardiorespiratory (CR) interactions. Moreover, the algorithm was designed, in such way that it also can be applied on (capacitively-coupled) ECG measures combined with an unobtrusive respiratory measurement. A journal paper was prepared based on these results and submitted to Physiological Measurement [38].

Chapter 8

In this chapter, the use of PPG features was investigated to classify 1-minute windows as apneic or normal sleep. PPG features known to be modulated by both respiration and the autonomous nervous system were included. These results were presented during the computing-in-cardiology conference in 2018 [39].

Chapter 9

This chapter presents multimodal algorithms for the detection of sleep apnea. First, the inclusion of the thoracic belt respiratory effort signal in the ECG algorithm was investigated. Next, the added value of adding PPG or ECG features to the SpO_2-based model was explored. Moreover, the use of a hidden Markov model to better characterize the temporal dependencies and to separate a wake class was investigated. Finally, the AHI estimation and OSA screening performances were compared for the different modality combinations and the influence of sleep scoring on the different detection algorithms is discussed.

Chapter 10

This chapter presents another important contribution of this work, as here, true unobtrusive OSA detection solutions were tested. The algorithms developed based on PSG signals in Chapter 7 and 8 were tested on recordings of unobtrusive sensors. The ECG-based algorithm was tested on ccECG signals combined with either the EDR estimate or the ccBioz. The PPG algorithm was tested on the Byteflies PPG patch recordings.

Part IV: Cardiovascular Comorbidities in Sleep Apnea

Chapter 11

The goal of chapter eleven is to look beyond the AHI and to better phenotype OSA patients for treatment prioritization. To this end, the presence of cardiovascular comorbidities was considered to differentiate between OSA patients. It was investigated, whether physiological markers extracted from the pulse oximetry can be linked to the CV status of the patient. Several SpO_2 and PPG markers were investigated and a LS-SVM classifier was built using the most predictive markers. In the future, these novel markers could be used to identify OSA patients at the highest risk of developing CV comorbidities sooner, so that treatment might prevent this development. Intermediate results of this research were presented in [42], and the final results were published in [36].

Chapter 12

Using the novel SpO_2 markers, which were selected in the previous chapter as most predictive for the CV status, an interpretable classification model was trained using mICS to assess the cardiovascular status of OSA subjects. Moreover, extra statistical tests were applied to gain more insights in the results. The model was also tested on two additional data sets in order to assess the parameters on healthy volunteers and to validate if it is possible to predict the cardiovascular status. The presented results were published in the IEEE TBME [37].

Chapter 13

The final chapter of the manuscript gives an overview of the main contributions of the research presented in this PhD thesis. A general discussion is provided and future research opportunities are presented.

1.5 Collaborations

The research presented in this PhD thesis was conducted in the BIOMED research group for biomedical data processing led by professor Sabine Van Huffel, as part of the STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics in the Department of Electrical Engineering of KU Leuven.

Collaborations with several other people from both within and outside the BIOMED group were set up in order to enhance the research.

The work was performed in close collaboration with professor Carolina Varon. In our weekly meetings she provided guidance and feedback to steer the PhD in the right direction. She also provided several algorithms that were used in this work, including the phase rectified signal averaging (PRSA) computation [10], EDR estimation [132] and the ECG-based apnea detection algorithm from [135], which was used as a basis for Chapter 7.

Dr. Jesús Lazáro shared his knowledge on PPG processing and shared his algorithms for PPG pre-processing. We collaborated on a first PPG-based apnea detection algorithm [78], on which the further work in Chapter 8 is built. There was also a nice collaboration with Dorien Huysmans, with whom I often shared thoughts about the research and data sets. The collaboration with John Morales started when he was a master thesis student under the supervision of Carolina and myself, which resulted in a conference paper [93]. In his PhD, he focuses on studying parameters which characterize cardiorespiratory signal interactions. He provided the algorithm to extract several CV interaction parameters used in Chapter 7. Jonathan Moeyersons also provided the ECG quality indicator used in that Chapter.

During his visit to the BIOMED group, I also collaborated with Javier Milagro from the University of Zaragoza on his study into heart rate variability (HRV) changes in OSA patients with CV comorbidities, compared to patients without any CV problems [89].

On the clinical side, the research was conducted in strong collaboration with the sleep laboratory of the University hospitals Leuven. Prof. Dries Testelmans, Prof. Bertien Buyse and Pascal Borzée provided most of the data sets used in this work. In our monthly meetings they also shared their research ideas and gave clinical insights into the obtained results.

Moreover, the PhD was funded as part of the VLAIO project (no. 150466) OSA+, a collaboration between KU Leuven, UZ Leuven, TU Eindhoven, Sleep medicine center Kempenhaeghe, and several industrial partners. Within this project, a data set of PSGs from healthy volunteers, recorded at the sleep center in Kempenhaeghe was shared. This data was included as a control set in our study into CV comorbidities in Chapter 12. Additionally, this PhD research received funding from the Flemish Government (AI Research Program).

For the collection of ccECG and ccBioz data, used in Chapter 10, we had a collaboration with Imec. Dr. Ivan Castro helped in the study design, provided the wearable, gave technical assistance during the measurements and helped with the interpretation of the obtained results.

Chapter 2

Data collection and labeling

This chapter will give an overview of the data that will be used in the remainder of the manuscript. The first section will provide some background on the acquisition principles for each signal modality. Next, more information is provided on the labeling of the cardiovascular status of OSA subjects. Finally, an overview will be given of the data sets that were analyzed during this PhD. These data sets were structured and pre-processed, to be ready-to-use for collaborators and future generations of students. The construction of these data sets was a main contribution of this PhD.

2.1 Description of the studied signal modalities

This section will discuss the acquisition principles of the oximetry, ECG-based and respiratory signals.

2.1.1 Oximetry: SpO_2 and PPG

Acquisition principles

Oximetry is a noninvasive optical measurement technique detecting blood volume changes, which was first applied in 1935 [127]. The following years, the acquisition principles and sensor design were optimized and by the mid 1980's, the technique was used in daily clinical practice to monitor the blood oxygen saturation and heart rate (HR).

The oximetry sensor consists of a light emitting diode (LED), which illuminates the skin, and a photodetector. Oximeters can be either of the transmissive or the reflective mode. In the case of a transmissive oximeter, the LED and photodetector are separated by the measured surface, and the photodetector captures the non-absorbed light that is transmitted through the surface. This mode is often used on fingers or earlobes. In the case of a reflective oximeter, the photodetector is placed at the same side as the LED and it captures the light that gets reflected by the surface [124]. Reflective oximeters are often applied to the wrist or the forehead.

For both types of oximeters, the amount of light detected by the photodetector is inversely related to the light absorbed by the measured surface. The light is absorbed by different components as depicted in Figure 2.1. The direct current (DC) components, on the one hand, vary slowly over time. These include absorption by the skin and the surrounding soft tissues, the venous blood and the non-pulsatile component of the arterial blood. The DC component can vary due to respiration, sympathetic system activity, and thermoregulation [124]. The alternating current (AC) component, on the other hand, pulsates according to the blood volume changes in the arteries caused by each heartbeat. During the systolic phase, the blood is pumped out of the heart, causing an increase in blood volume resulting in an increase in light absorption. During, the diastolic filling phase of the heart, the opposite is observed. The resulting signal that is

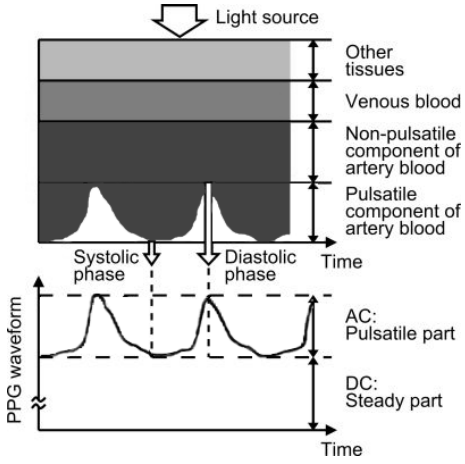


Figure 2.1: Overview of the working principles of an oximeter. The absorption diagram of the light emitted by the oximeter is given on top. At the bottom, the resulting waveform captured by the photodetector is shown. This waveform can be divided in a steady and a pulsatile part. Figure adapted from [124].

obtained by the photodetector, is often called the pulse photoplethysmogram. The cardiac cycle and HR can be extracted from the PPG, but the signal has also been studied to estimate the blood pressure, cardiac output and respiration, as well as to detect arterial diseases [3].

Adult blood is mainly composed of oxyhemoglobin (O_2Hb) and reduced hemoglobin (Hb). Both have a different light absorbance, depending on the wavelength of the emitted light, as depicted in Figure 2.2, where the light extinction coefficient is plotted as measure of absorbance. By analyzing the amount of light absorbed when different frequencies of light are emitted, the oxygenation level of the arterial blood can be defined. Red (660 nm) and infrared (IR, 940 nm) light are commonly applied in this case, since at these wavelengths a clear difference between the O_2Hb and Hb curve can be observed. The SpO_2 signal is estimated using empirically defined calibration curves on the ratio (R) of the AC and DC components of the oximetry signal [127]:

$$R = \frac{AC_{red}/DC_{red}}{AC_{IR}/DC_{IR}}. \tag{2.1}$$

Clinical oximetry

During PSG recordings, a transmissive oximeter on the fingertip is commonly applied. At UZ Leuven, a Nonin 8000J sensor is used. This sensor makes use of red and IR light to record the SpO_2 and PPG signals.

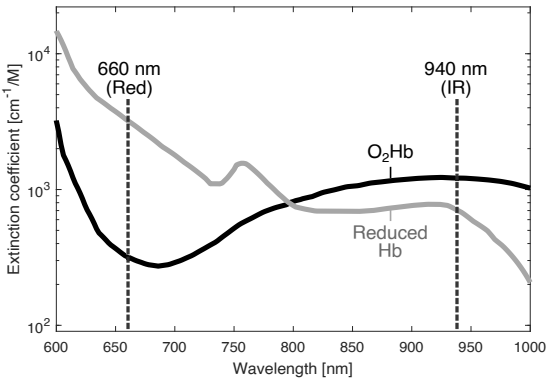


Figure 2.2: Light absorbance spectra for O_2Hb and Hb in water. Figure generated using data from [99].

Byteflies PPG sensor dot

Within the OSA+ project, wearable PPG signals were acquired simultaneously with the gold-standard PSG. For this, we used the Byteflies sensor dot, which is depicted in Figure 2.3. The dots were placed at the wrist to simulate a wristband-like device and collect reflective PPG signals at a sampling rate of 100 Hz for three different light frequencies: red (660 nm), IR (940 nm) and green (535 nm). The dot also includes 3-axis accelerometry (ACM).



Figure 2.3: Byteflies sensor dot.

2.1.2 Electrocardiography signals

Electrocardiogram

The ECG measures the electrical activity of the heart by placing surface electrodes on the skin. The characteristic waveform of an ECG is presented in Figure 2.4, two heartbeats are shown. Each heartbeat starts with an electrical pulse fired by the sino-atrial node, which acts as an internal pacemaker. This pulse causes a slow depolarization and contraction of the atria, which is captured by the P-wave. The electrical pulse is then delayed by the atrio-ventricular node, such that the transfer of blood from the atria to the ventricles can be completed. Next, the His bundle and Purkinje system conduct the pulse to the ventricles where a rapid contraction takes place, pumping the blood into the arteries. This contraction causes the most prominent peak in the ECG signal, the QRS-complex. After which, an iso-electric segment known as the ST segment occurs due to the long action potential duration of the ventricular muscles. Finally, the heartbeat ends with the relaxation of the ventricles causing the slow T wave [110].

Cardiologists use the ECG to assess the cardiovascular state of patients, since the ECG waveform is altered by cardiovascular diseases and abnormalities. To have a complete characterization of the ECG waveform, a 12-lead ECG is typically used in this case, measured using 6 electrodes placed on the chest and 4 on the limbs [110]. In patient monitoring systems, one is mainly interested in

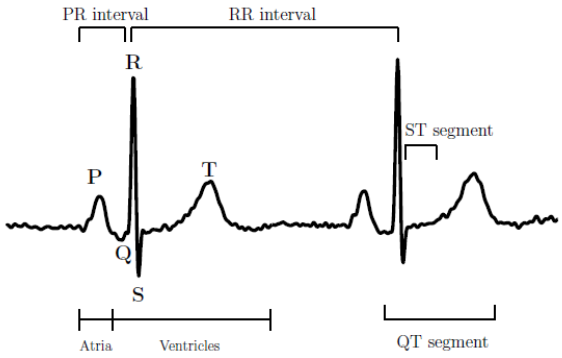


Figure 2.4: Characteristic waves of a normal ECG. Figure taken from [129].

the HR, which is measured using the time difference between two consecutive R-peaks, called the RR-interval. For this application a 1-lead measurement usually suffices. Often lead II is chosen, which places one electrode on the right arm and the other one on the left leg. In the case of a PSG, a modified lead II is typically applied, placing the electrodes on the torso instead [136].

The HR will be controlled by the autonomic nervous system, which is subdivided in a sympathetic and a parasympathetic branch. The sympathetic nervous system (SNS) reacts to stressful situations (fight-or-flight), by increasing the HR. When an apnea occurs, the SNS will kick in causing the arousal, thereby restoring the breathing. The parasympathetic nervous system (PNS), on the other hand, will cause the state of rest-and-digest. It will help in the recovery of stressful events by slowing down the HR and respiration [129]. The interactions of the SNS and PNS are studied by investigating the HRV in the frequency domain. Mainly the power in the low-frequency range (LF, 0.04-0.15 Hz) and the high frequency range (HF, 0.15-0.4 Hz) have proven to be relevant. The HF band is often linked to the modulation of the HR by the respiration, called respiratory sinus arrhythmia (RSA), which is controlled by the PNS. The LF band, on the other hand, is controlled by both the SNS and PNS [16].

The respiration will not only affect the HR, but will also have an influence on the ECG morphology. This is due to the mechanical effects of the respiration on the ECG: the respiratory movements of the thorax will cause changes in the electrode position, and the electrical impedance of the thorax will vary with respect to the air volume in the lungs. The morphological changes caused by these effects are captured by EDR algorithms to estimate the respiration from the ECG [132].

Capacitively-coupled ECG

ECG measurements using electrodes attached to the skin, are not ideal for long-term measurements as they can cause skin irritations. Therefore, non-contact ECG systems have been developed. Within this PhD, ccECG measurements were recorded simultaneously to the gold-standard PSG, using a sensor developed by imec, which was integrated in a mattress cover. In this ccECG sensor, the galvanic contact between the skin and the electrode is replaced by a capacitive coupling between the skin and the conductive plate of the sensor. All non-conductive materials in between, such as the bed sheets and clothing, serve as dielectric. The sensor includes an array of electrodes, as can be seen from the recording setup in Figure 2.5. A quality-based electrode scanning algorithm selects the optimal electrode configuration every minute. As such, the sensor adapts its electrode configuration when, for example, the sleeping position of the patient changes. More information on the sensor setup and electrode scanning algorithm can be found in [25] and [26].

2.1.3 Respiratory signals

Thoracic belt

As mentioned in Chapter 1, the PSG includes different measures of respiration. In this research, the thoracic respiratory effort belt (TB) was used, as this sensor is less obtrusive than the oronasal thermistor and nasal pressure sensor. Moreover, the thoracic belt provides more stable measurements than the

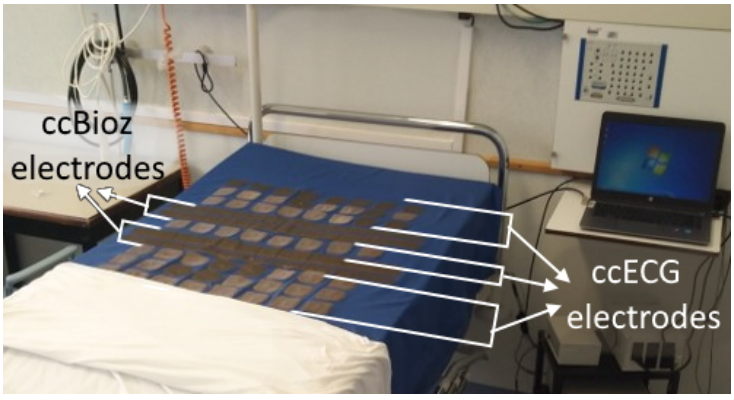


Figure 2.5: Recording setup of the mattress with non-contact sensors embedded, measuring ccECG and ccBioz.

abdominal one, since in patients with a high BMI the position of the abdominal sensor often shifts throughout the night. The TB measures the respiration using the RIP technology. The sensor includes wire coils which generate an electromagnetic field. When the diameter of the coils varies due to respiratory movements, the recorded electromagnetic field will change accordingly [136].

Capacitively-coupled bioimpedance

In order to make the respiratory measurement even more unobtrusive, the use of a non-contact ccBioz sensor was investigated. This sensor was included in the same mattress cover as the ccECG, as shown in Figure 2.5. The ccBioz sensor injects a high frequency current through the subject's chest and measures the resulting voltage. Changes in air volume in the lungs will cause variations in the bioimpedance of the chest and thus also in the measured voltage [27].

2.2 Cardiovascular labels

For most of the data sets, the cardiovascular status of each patient was assessed at the time of the PSG. The presence of apnea-associated comorbidities like hypertension, hyperlipidemia, diabetes, atrial fibrillation and peripheral arterial disease was determined. This was done based on either the presence of the condition in the history of the patient, the patient's drug use, or the detection of the condition by means of blood pressure and blood analysis tests. All these conditions were grouped in the cardiovascular "condition" (cC) category. The patient history was also examined for the occurrence of more severe cardiovascular "events" (cE), including myocardial infarction, heart failure and stroke. Subjects for whom none of the mentioned cardiovascular conditions were observed were considered as "controls" (Ct).

2.3 Overview of the used data sets

A general overview of the used data sets can be found in Table 2.1. For each data set, the table presents the number of recordings, patient demographics, average AHI and the percentage of recordings with an AHI larger or equal than 15. It is also depicted whether information on the cardiovascular status is available, and additional information is presented in the last column. The data sets will be described in more detail below. They are grouped according to their origin, starting with the data sets recorded at the sleep laboratory

of the University hospitals Leuven, followed by the publicly available Sleep Heart Health Study (SHHS) and Apnea-ECG data sets, and ending with the Kempenhaeghe set of controls.

2.3.1 Data sets recorded at UZ Leuven

These data sets include PSGs recorded at the sleep laboratory of the University Hospitals Leuven, from patients who were suspected to suffer from sleep apnea. All signal modalities were recorded with a sampling frequency of 500 Hz, although for some data sets only downsampled versions of the recordings were available as specified in the data set description. The PSG data was manually annotated by sleep experts, the sleep stages were scored, and apneic events were annotated according to the AASM 2012 scoring rules [17], except for the "Varon" and "Cluster" data sets, which were scored using the Chicago criteria [107]. The collection of these data sets was approved by the ethical committee of UZ Leuven (S53746, S60319) and all patients signed an informed consent. Specific information on each data set will be given in the next paragraphs.

Matched

The "Matched" data set was compiled by selecting patients with an AHI larger than 15 from the hospital's database. Subjects with a CV condition or event were matched for age, gender, body mass index (BMI) and smoking status with a CV control patient. A data set of 100 patients was collected, including 78 men and 22 women. Hyperlipidaemia and hypertension were the most common comorbidities, present in, respectively, 46 and 40 subjects. 5 other patients suffered from diabetes, 4 patients had a myocardial infarction and 2 a stroke.

OSA+

The "OSA+" data set is a prospective data set recorded within the OSA+ project. During these PSG recordings, two wearable measurements were simultaneously performed. A Byteflies PPG sensor dot was placed on the wrist measuring wearable PPG, and an EMFIT pressure sensor was placed on the mattress. In this thesis, only the PPG data was considered, the EMFIT processing was outside the scope of this research. In the first recording phase of 102 subjects, the PPG sensor dot was not yet applied. From the following 116 recordings, the Byteflies PPG was successfully recorded for 100 of them.

Table 2.1: Overview of the used data sets including the number of recordings (# Rec.) per data set, the patient demographics, AHI and the percentage of recordings that had an AHI equal or larger than 15. The availability of information on the cardiovascular status of the subjects is also depicted and additional information was added in the last column.

		# Rec.	Age Years	BMI Kg/m^2	Male	AHI Events/h	AHI ≥15	CV info	Additional Info
UZ Leuven	Matched	100	48±11 (38, 56)	30±5 (27, 33)	78 %	41±22 (24, 57)	100 %	yes	CV status 1-to-1 matched
	OSA+	218	49±12 (41, 58)	30±6 (25, 34)	65 %	27±24 (8, 39)	58 %	yes	Byteflies PPG for 100
	Imec	28	48±14 (38, 55)	30±5 (25, 35)	50 %	28±26 (6, 58)	57 %	no	ccECG and ccBioz
	Cluster	677	52±12 (45, 61)	31±6 (27, 35)	80 %	40±20 (23, 53)	100 %	yes	Downsampled signals
	Varon	10	48±11 (43, 54)	30±5 (27, 31)	90 %	33±20 (27, 44)	80 %	no	Chicago criteria
SHHS	SHHS1	5793	63±11 (55, 72)	28±5 (25, 31)	53 %	18±16 (7, 23)	45 %	yes	General population
	SHHS2	2651	67±10 (60, 76)	28±5 (25, 31)	54 %	18±16 (7, 23)	45 %	yes	
Physionet Apnea-ECG	ECG	70	45±11 (38, 54)	28±6 (22, 33)	81 %	28±27 (0, 43)	60 %	no	1 min apnea labels
	SpO_2	8	43±8 (38, 52)	28±8 (22, 35)	88 %	32±36 (0, 70)	50 %	no	
Kempenhaeghe	Controls	83	34±13 (22, 46)	24±3 (22, 25)	33 %	1±1 (0, 2)	0 %	yes	Healthy volunteers

Age, BMI and AHI are presented as mean ± standard deviation, with the 25 % and 75% quantile values underneath.

Next to the addition of wearables, the "OSA+" data set also includes a cardiovascular assessment for 197 of the recordings. The other 21 subjects had missing blood values, making it impossible to rule out the presence of hyperlipidemia and diabetes. In total, 46 Ct, 140 cC and 11 cE patients were included. In contrast to the "Matched" data set, this set includes a full range of AHI values, 42 % of the subjects had no or only mild OSA.

Imec

The "Imec" data set was a follow-up of the "OSA+" data set, in which the EMFIT sensor was replaced with an array of sensors embedded in the mattress, as depicted in Figure 2.5, which measure ccECG and ccBioz. The data set consist of 28 recordings, but due to technical problems (i.e. unidentified electrode disconnection in the prototype), the ccBioz signal could only be collected for 11 subjects. Additionally, wearable PPG signals were still recorded using the Byteflies PPG sensor dot on the wrist.

Cluster

The "Cluster" data set was retrospectively composed from diagnostic PSGs recorded between 2009 and 2012 from patients who started CPAP therapy. The data set was composed by the sleep specialists of the UZ Leuven to study OSA-related comorbidities. In this research, only the cardiovascular comorbidities were studied. The PSG signals and a full CV assessment was available for 677 subjects, of which 72 were classified as Ct, 546 as cC, and 59 as cE.

As this data set was composed from archived recordings, the PSG signals were down sampled. About one third of them was available at 200 Hz, the other ones had a sampling frequency of 50 Hz. Moreover, the apnea labeling was performed using the older Chicago criteria, therefore this data set is not recommended to develop algorithms for the automatic detection of sleep apnea.

Varon

In order to compare the ECG based algorithms with the previous work by Varon et al. [135], the "Varon" data set was also included. This data set consists of 10 ECG recordings extracted from PSGs sampled at 200 Hz and scored for apneic events according to the Chicago criteria.

2.3.2 Sleep heart health study

The SHHS is a multi-center cohort study focusing on the cardiovascular and other consequences of sleep-disordered breathing, which is made available by the National Sleep Research Resource (NSRR) [34, 108, 111]. Recordings from 5793 persons from the general population undergoing unattended full night PSG at baseline are available. All subjects were 40 years or older and had no history of treatment for OSA. 2651 of these subjects underwent a follow-up PSG, on average 5 years later. Moreover, their cardiovascular disease outcomes were monitored for up to 15 years. As can be observed in Table 2.1, these subjects are with an average age of 63 years at baseline significantly older than the clinical populations recorded at UZ Leuven, which have an average age of about 50 years. Moreover, the BMI is slightly lower, more women are included, and the average AHI is much lower in the SHHS sets compared to the clinical data sets. As these subjects were selected from the general population, most of them did not suffer from any apnea-related symptoms either.

Full PSG recordings are available, but no PPG measurements were performed. The recordings were manually scored; annotations of sleep stages, arousals, oxygen desaturations and respiratory events are available. This enables the user to apply different hypopnea definitions. In this research, the AHI was defined taking into account all apneas and hypopneas with either an arousal or an oxygen desaturation of at least 3 %, in order to match the AASM 2012 rules [17]. This AHI is available as the "ahi_a0h3a" in the original polysomnography variables included in the data set. Annotations of individual apneic events were also extracted, ignoring labeled hypopnea annotations without either an arousal or desaturation of at least 3 %. It is, however, a known issue with the data set that original oxygen desaturation annotations were indelibly changed due to software conversions over the years. Therefore, these annotations were not used, the oxygen desaturations were automatically recomputed and linked to hypopneas, as described in Chapter 6. Based on these linked desaturations, it was decided which hypopneas should be taken into account.

2.3.3 Physionet Apnea-ECG

The Apnea-ECG data set [104] can be found on Physionet [54]. It consists of 70 full night ECG recordings, of which 8 also include the SpO_2 and respiratory signals. The AHI and minute-by-minute apnea annotations are available, depicting if any apneic event occurred during each 1-minute segment. 31 of the recordings had severe OSA and 23 were recorded from healthy volunteers having AHIs of less than 1. The over-representation of those two extremes might

explain why OSA screening algorithms generally obtain better performances on this data set, than on clinical data sets [101].

2.3.4 Kempenhaeghe controls

The last data set was recorded at the Kempenhaeghe Center for Sleep Medicine in Heeze as part of the Healthbed study. It consists of PSG recordings of 83 healthy volunteers, that did not have any sleep disorders, nor any cardiovascular comorbidity. All recordings had an AHI lower than 5. The collection of this data set was approved by the Kempenhaeghe ethical committee and by the medical ethical committee of the Maxima Medical Center Eindhoven (W17.128).

Chapter 3

Literature review: Detection, screening and phenotyping of OSA

Many researchers have been developing methods for automated in-home screening and diagnosis of sleep apnea as an alternative for the in-hospital PSG. In this chapter, an overview will be given of methods using signals that can be easily acquired in a home environment. Different methods will be introduced, based on, cardiovascular measures, oximetry, effort and respiration, and sound. Thereafter, an overview of deep learning methods will be given, followed by multimodal detection algorithms, and a comparison of methods and modalities. Next, an overview of currently available OSA screening devices will be provided. The chapter will be concluded with a section on phenotyping OSA beyond the AHI.

This chapter is included in the book chapter ***Deviaene, M.*** *, Huysmans, D., Castro, I., Borzée, P., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Sleep monitoring in adults using wearables and unobtrusive technology, which was submitted for publication in "Wearable/Personal monitoring devices present to future", G. Gargiulo and G. R. Naik, Springer, 2020 [35].*

3.1 Introduction on OSA screening

The discussed OSA screening methods will be ordered according to the used measurement modality, including Sleep, Cardiovascular, Oximetry, Position, Effort and Respiratory parameters (SCOPER). This classification system for portable sleep monitoring devices, was introduced by Collop et al. [30]. A short overview of the use of these five categories of sensors in the diagnosis of sleep apnea is given below.

1. **Sleep:** A measurement of sleep, for example using actigraphy, will enable the calculation of the hours of sleep. Using this measure instead of the recording time leads to a better estimation of the AHI [96]. These methods are outside the scope of this thesis and will not be discussed further.

2. **Cardiovascular:** This category includes ECG, PPG and peripheral arterial tonometry (PAT) as well as all other measures of the HR. Bradycardia can be observed during apneas, followed by tachycardia when breathing is retaken [57], as can be observed in Figure 3.1. Moreover, these signals can be used to derive an estimate of the respiration and detect autonomic arousals.

3. **Oximetry:** The SpO_2 signal is very useful for OSA screening, since apneas often result in an oxygen desaturation as can be seen from the example in Figure 3.1.

4. **Position:** Visual and non-visual measures (e.g. using an accelerometer combined with a gyroscope [69]) can determine the body position of a subject. Studies have shown that OSA severity can be position dependent. In fact, a higher AHI and more severe apneas have been observed during supine sleep [100]. Differentiating between lateral and supine sleep could thus help to improve automated OSA detection algorithms. No studies in the literature, however, propose to use the position measurement on its own for the detection of OSA. Therefore, the analysis of position sensors will not be discussed further.

5. **Effort and Respiration:** The reduction in airflow is the primary effect of apneas. Sensors measuring respiratory effort provide extra information which helps to differentiate between central and obstructive apneas.

In addition, in this literature review, sound recording devices were defined as a sixth category. In this chapter, the focus will be on the state-of-the-art detection algorithms, rather than on the wearable sensors. An overview of current research and commercial devices for OSA detection categorized according to the SCOPER

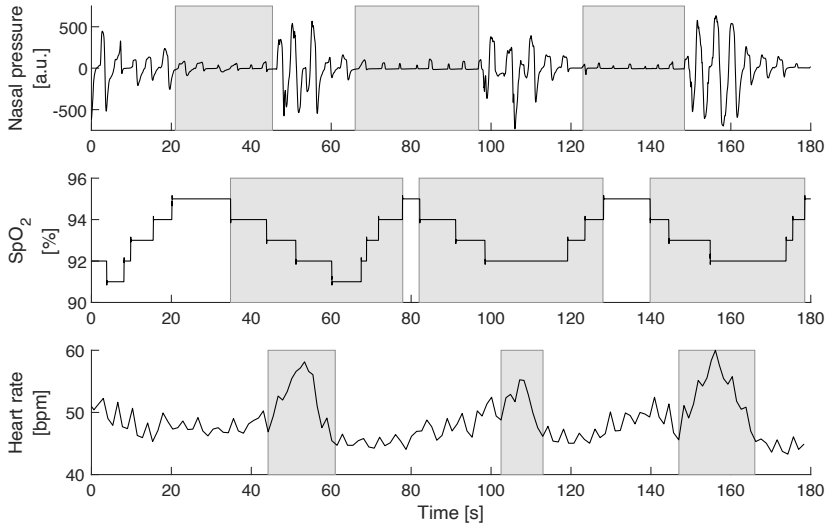


Figure 3.1: Example of a segment with apneas. From top to bottom, the nasal pressure, SpO_2 and HR signals are plotted, with respectively, the annotated apneas, oxygen desaturations and sympathetic activation depicted by the shaded areas.

system can be found in [85]. In Section 3.9, a short discussion on the available OSA screening devices is provided.

Most state-of-the-art OSA detection algorithms consist of 4 main steps: starting with the signal preprocessing phase, next feature extraction, feature selection and finally classification. Two classification problems are typically studied: event-based or subject-based classification. In the case of event-based classification, the goal is to detect all apneas within the recording, this is often done by splitting the recording into 1-minute windows. Based on the number of windows classified as apneic, the AHI can be estimated. Subject-based classification, on the other hand, extracts features over the whole recording which will then be correlated to the AHI and used to predict the OSA severity category of a patient. In this section, features commonly used for either one of those classification tasks in the framework of OSA detection, will be discussed per SCOPER category. Additionally, studies describing automatically generated features using deep learning, will be briefly discussed in Section 3.6.

3.2 Cardiovascular

3.2.1 ECG

The effects of apneas on the ECG signal have been known since the 1980's [57, 92]. But the research into sleep apnea screening using ECG signals really took a boost in 2000 with the Computers-in-Cardiology Challenge [91], and the subsequent release of the Physionet Apnea-ECG database [104]. This data set is still one of the most used data sets for ECG-based OSA screening. An extensive overview of ECG-based methods for OSA screening can be found in [50, 86]. The ECG-based screening approaches proposed in [86] obtain subject-based classification accuracies of 72-100 % and an area under the reiceiver operating characteristic curve (AUC) of 89-100 %.

The used features are generally based either on changes in HR [57], or changes in respiration, captured using the EDR [92]. The respiration can be extracted from the ECG since ECG waveform characteristics are altered by respiratory-induced movements of the chest electrodes and changes of the electrical impedance of the thoracic cavity. The simplest measure of EDR is the amplitude of the R-waves, but many more elaborate methods have been proposed. A comparative study of 10 EDR algorithms is discussed in [132], where it was shown that the EDR signals computed from changes in the QRS slopes are most robust in the presence of artefacts, changing respiratory rates, and different ECG-recording systems.

OSA detection algorithms typically use statistical, frequency domain and nonlinear features extracted from both the HR and the EDR. Recently, measures of cardiovascular interactions have also been identified as relevant for OSA detection [135].

Statistical features: Different time-domain features of the RR-interval time series and EDR are often used in the literature [33, 82]. An overview of these features can be found in Table 3.1.

Frequency domain features: Multiple methods have been applied to extract frequency information from the HR, but the power spectral density and wavelet decomposition [22] are the most popular ones. Often the power in the standard HRV frequency bands is considered: the LF (0.04-0.15 Hz) and the HF (0.15-0.4 Hz) band.

Nonlinear features: An overview of nonlinear parameters used in cardiorespiratory signals for sleep apnea detection is given in [133]. Analysis of the sample entropy of the HRV signal, has shown a reduced complexity for apneic subjects compared to controls [2]. This can be linked to the cyclic

pattern of bradycardia and tachycardia during apneas. Detrended fluctuation analysis has also been applied on the HRV of apneic subjects, but only slight changes in long-term control mechanisms of the HR could be observed [133].

Cardiorespiratory interactions: The HRV is modulated by the respiration through the mechanism called respiratory sinus arrhythmia. It has been observed that the information shared between respiration and HRV decreases during apneas and these reductions have been used as features to detect apneas [135]. An overview of different methods to estimate the RSA from the HRV and a respiratory measurement (e.g. the EDR or a respiratory belt) can be found in [94].

Table 3.1: Overview of commonly used ECG time-domain features.

RR	
Mean	μ_{RR}
Standard deviation	σ_{RR}
Serial correlation coefficients (of order k)	$r_k = \frac{\sum_{i=1}^{m}(RR_i - \mu_{RR})(RR_{i+k} - \mu_{RR})}{\sum_{i=1}^{m}(RR_i - \mu_{RR})^2}$
NN50	# adjacent RR-intervals that differ more than 50 ms
pNN50	pNN50=$\frac{NN50}{m}$
Standard deviation of interbeat intervals	SDSD = σ_{rd}, with $rd_i = RR_{i+1}$
Root mean square of interbeat differentials	RMSSD = $\sqrt{E(rd_i^2)}$
Allan factor	A(T)=$\frac{E[(N_{i+1}(T)-N_i(T))^2]}{2E[N_{i+1}(T)]}$ with $N_i(T)$= # beats in the ith window of T seconds
EDR	
Mean	μ_{EDR}
Standard deviation	σ_{EDR}

3.2.2 PPG

From the PPG signal, the pulse rate (PR) can be extracted as alternative for the HR. All features proposed for the HR could therefore also be estimated from the PR. Moreover, decreases in the amplitude fluctuation of the PPG (DAP) have been linked to vasoconstriction due to sympathetic arousals at the end of apneas. The use of these DAP events has been studied for the detection of OSA in children [53]. Additionally, characteristics of the PPG waveform such as pulse amplitude, width and slope transit time have been linked to respiration and their behavior can be useful for sleep apnea detection [78, 102].

3.2.3 PAT

The peripheral arterial tonometry [117] sensor was developed for the detection of sleep apnea. It is based on a pulsatile plethysmography signal measured on the finger which can capture vasoconstriction and tachycardia related to apneas. Decreases in PAT amplitude of 33 % in combination with a tachycardia of 15 % were used in [117] for the detection of apneas.

3.2.4 Contactless cardiovascular measurements

Several contactless alternatives for ECG recording have been tested for sleep apnea detection. In [25], a multi-channel contactless capacitively-coupled electrocardiography embedded in a mattress was proposed. The ECG-based apnea features defined in [135] were extracted from ccECG signals of healthy volunteers. The HR features achieved high similarity to the features extracted from the reference ECG signal. Signal morphology features, on the other hand showed lower similarity with the reference resulting in problems with the EDR extraction. The similarity in HR features shows the potential of ccECG sensors for the detection of sleep apnea.

Ballistocardiography (BCG) has also been studied in sleep apnea subjects. In [144], for instance, Zink et al. optimized the algorithm for BCG beat-to-beat detection in a sleep apnea population and obtained an R^2 correlation coefficient of 0.95 between the beat-to-beat cycle lengths extracted from the BCG and ECG. Whereas Hwang et al. investigated the correlation between the oxygen desaturation index (ODI, see section 3.3) and HR features extracted from three different BCG sensors [63].

Researchers have also focused on extracting HR from radar or IR video recordings next to the breathing movements. Zhu et al. have validated IR-based HR

monitoring in sleep apnea subjects and discussed the open challenges in this field [142].

3.3 Oximetry

The SpO_2 signal has been widely studied for sleep apnea screening since it can easily be acquired in a home environment. In clinical practice the SpO_2 signal is often characterized by simple statistics such as the ODI, which counts the number of times a desaturation larger than 4 % occurs per hour of sleep, or time spent below an oxygenation level of 90 % (T90). It has been shown that these simple parameters can reliably confirm moderate to severe OSA, but that a negative test result does not rule out mild OSA [126]. Therefore, researchers have been developing a range of more elaborate SpO_2 parameters in order to detect subtle changes due to apneas. In [126], four categories of commonly used SpO_2 parameters were defined: desaturation characteristics, time series variables, frequency spectrum variables and nonlinear variables.

Desaturation characteristics: The desaturation characteristics include the desaturation depth, duration and area. These can be computed compared to a baseline level, or another reference point [138]. Additionally, these parameters can be calculated on subparts of the desaturation event, e.g. on the downward and upward part separately. Some studies also considered desaturation slopes, or higher order derivatives of the signals [72]. In Figure 3.2 an overview of these features is given for an example oxygen desaturation.

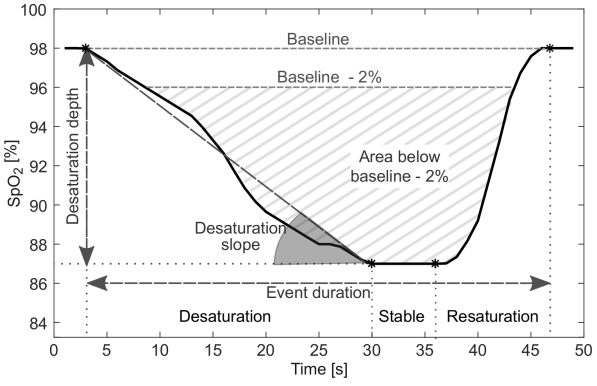


Figure 3.2: Overview of desaturation characteristics which can be extracted from oxygen desaturations.

Time series analysis variables: These parameters include statistical measures that can be computed over the full SpO_2 signal, or window-based: the mean, minimum, quantile values, standard deviation, higher order statistical moments, and cumulative time spent below certain SpO_2 values (e.g. T90) [126].

Frequency spectrum variables: Studies have shown that OSA subjects have an unstable or overly sensitive respiratory control system [44]. Therefore, instead of going back to a stable breathing pattern after an apnea, often another apnea is triggered. As such, apneas tend to be grouped together, as can be seen from Figure 3.3. When the apneas follow each other closely, a periodic pattern of desaturations can occur as shown in the lower right plot of Figure 3.3. Different studies have used the power in the 0.01-0.033 Hz frequency band to capture this periodicity [126]. In [93], the use of wavelets was proposed instead.

Nonlinear analysis variables: The recurrent desaturations during the night also cause the SpO_2 signals to have a higher irregularity, variability and complexity. These can, respectively, be measured using the sample entropy, central tendency measure and Lempel-Ziv complexity [4].

Terrill et al. concluded in their review [126] that the studies which obtained the best OSA detection results used features from multiple of these four categories. These categories thus contain complementary information. However, including more than three to four features in the classifier did not lead to a significant

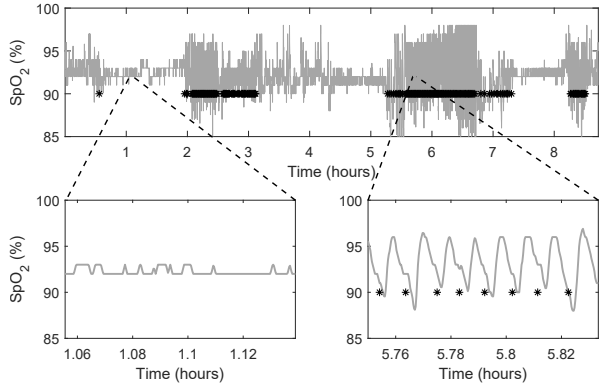


Figure 3.3: Example SpO_2 signal of a subject with an AHI of 39.4 in which the apneas (*) are grouped together. In the lower right plot a 5 min apneic segment is shown in which periodicity can be observed in the SpO_2 signal, whereas the left plot shows no periodicity during normal sleep.

.

increase in performance. SpO_2-based OSA screening methods typically obtain accuracies of 80-95 % with an AUC of 90-95 %, depending on the used features, data set and AHI threshold [126].

3.4 Effort and Respiration

3.4.1 Airflow

The AASM rules for visual scoring of apneic events rely mainly on the airflow measured using an oronasal thermal and nasal pressure sensor [17]. Therefore, several researchers have developed methods for the automatic detection of sleep apnea on these signals. All drops in peak signal amplitude of more than 90 % are scored as apneas, whereas drops of more than 30 % could represent a hypopnea. These drops are relative to pre-event baseline breathing which is defined as "the mean amplitude of stable breathing and oxygenation in the 2 minutes preceding the onset of the event or the mean amplitude of the 3 largest breaths in the 2 minutes preceding the onset of the event" [17]. This baseline can be difficult to define for subjects with high AHIs and its definition is ambiguous to be translated into algorithms. Moreover, breathing and movement artefacts will further complicate this baseline extraction for automated algorithms [29]. Therefore, Cioleck et al. proposed a robust airflow envelope tracking using nonlinear filters. Using this reliable baseline, rule-based methods can be applied to detect apneas and hypopneas when a decrease in airflow amplitude larger than a predefined threshold is observed.

These rule-based methods, however, have as disadvantage that they cannot differentiate whether hypopneas are associated with arousals or desaturations. Therefore, more elaborate respiratory features have been proposed in order to capture changes that might have a correlation with arousals or desaturations. Koley et al. proposed statistical features of the respiratory amplitudes and intervals over windows of 8 seconds [71]. Additionally, they extracted the area and length of the respiratory curve and frequency domain features using the power spectral density (PSD). In this study, the deviation of each of the feature values from the subject's mean values were also considered.

3.4.2 Respiratory effort

Similar processing methods can be applied to respiratory effort signals obtained by RIP abdominal and thoracic belts. Despite the absence of airflow during obstructive apneas, the respiratory effort is still present, which challenges the

detection of apneas when using this signal modality. However, when the effort is measured in combination with other modalities like airflow or SpO_2, it can be used to differentiate between central and obstructive events [95]. When both the abdominal and thoracic belts are monitored, the thoracoabdominal paradox can also be used as marker for apneic events [17]. This paradox represents the phenomenon in which the respiratory effort signals from the thoracic and abdominal belts are out of phase. This is often observed during obstructive apneas.

A more extensive overview of methods using airflow and respiratory effort signals to detect OSA can be found in [86, 128]. The methods based on respiration reported in [86] obtained subject-based classification accuracies of 82-96 % and an AUC of 88-90 %.

3.4.3 Contactless measures of respiration

A lot of research has been conducted towards contactless sensors to measure respiratory effort. Respiratory movements can be extracted by pressure sensors in the mattress. An example of such sensor is the Emfit. The use of this sensor for the detection of sleep apnea was investigated in [62]. Rather than focusing on the breathing cessations, this study used unsupervised learning (i.e., clustering) to detect breathing distortions in the respiratory signal which are linked to arousals and large chest movements when breathing is regained. The amount of breathing distortions could be correlated to the AHI.

Kagawe et al. [67], proposed a system in which two Doppler radars are installed underneath the mattress in order to measure thoracic and abdominal effort. The system was tested for the detection of sleep apnea. A rule-based algorithm including the detection of paradoxical breathing showed promising results.

In [142], it was shown that the breathing rate can also be extracted accurately from IR video recordings in OSA subjects.

Castro et al. [26], proposed to measure the respiration using a capacitively-coupled bioimpedance sensor embedded in a mattress. This system, however, still needed to be tested on OSA subjects.

3.5 Sound

Snoring is one of the most common symptoms of OSA and differences in acoustic characteristics of snoring have been observed between OSA patients and simple

snorers [119]. Moreover, due to apneas, subjects often make grasping and choking sounds. Researchers have developed acoustic features which can be used for the automatic detection of sleep apnea using sound recordings. Solà-Soler et al. have proposed snore intensity, pitch, frequency, and spectral envelope parameters [119]. Their study showed that the snore-to-snore variability of these features was increased in OSA subjects. Moreover, multiscale entropy [11] of audio recordings has also shown to be useful to detect apneas.

3.6 Automatically generated features using deep learning

All features mentioned before are handcrafted, they are extracted based on previous knowledge of the effect apneas have on the signals. The signals might, however, contain more information than these features. Therefore, more and more researchers have moved towards automatic feature extraction using deep learning, avoiding the challenging task of feature engineering. In [97], 21 studies detecting sleep apnea using deep learning were compared. Most studies were based on the ECG signal, but also SpO_2 and respiration were considered. The convolutional neural network (CNN) was the most popular network. It is, however, not clear whether this is the best choice, since some studies suggested that recurrent neural networks (RNN) outperformed CNN [97]. The included ECG-based methods report global screening accuracies of 79.5-100 %.

3.7 Multimodal OSA detection

Since the scoring rules for sleep apnea are based both on a reduction in airflow and the presence of an oxygen desaturation, the combination of respiration and oximetry has frequently been studied [95, 128]. Moreover, cardiac signals are often added to capture the presence of arousals. Xie et al. [140] investigated the combination of SpO_2 and ECG-based features. They concluded that the SpO_2-based classifier outperformed the ECG-based one, and that only a slight increase in performance could be obtained when both feature sets were combined. Behar et al. [11], on the other hand, tested the combination of oximetry, sound and actigraphy. When comparing the different modalities, oximetry outperformed the other two with a classification accuracy of 85.1 %, but an increase in performance of 3.3 % was obtained when all three modalities were combined.

3.8 Comparison of methods and modalities

When comparing the OSA detection performances which are reported by studies using different modalities, automatic classification using respiratory signals obtained lower performances than ECG and SpO_2-based methods. Although, manual scoring of OSA is mainly based on the respiration. This might be due to a higher noise level in the respiratory signals [86].

ECG-based methods seem to outperform other modalities in single sensor applications. But this might be due to the fact that many ECG-based algorithms use public data sets, such as the Apnea-ECG database, which are less distorted by noise and might be easier to classify [86]. When different modalities are compared on the same data set, the SpO_2 signal performs best most of the times [140]. In order to validate the potential added value of deep learning compared to feature-based algorithms, comparative studies on more challenging datasets than the Apnea-ECG data set should be performed.

Another observation is the fact that combining different sensor modalities often does not result in a relevant improvement of performance, probably because one of the sensors performs quite well and dominates the analysis [86]. When SpO_2-based methods are combined with other modalities, they tend to dominate the performance [140]. Not all apneic events can, however, be detected using the SpO_2 signal, since the AASM rules allow scoring of apneas when no desaturation is present and hypopneas with only an arousal [17]. Therefore, sometimes the SpO_2 remains stable during apneas. Further research should be conducted to better merge different modalities in order to detect apneas which were missed by individual modalities. Additionally, combining different modalities improves the classification of apneas into event subtypes (central vs. obstructive) [95].

Although the Apnea-ECG database has stimulated the research into apnea detection methods greatly, most of the recordings only contain ECG signals. The database can therefore not be used to develop multimodal algorithms, nor can it be used to compare the performances of different modalities. Moreover, the control group was composed of healthy subjects, instead of subjects with other sleep related complaints which are seen in the sleep lab. Papini et al. have shown in their study [101], a drop in accuracy of 15 % and a drop in sensitivity of 30 % for algorithms trained on the apnea-ECG data set which were tested on a database including a wider spectrum of sleep disorders. These results show that the performance of apnea detection algorithms is strongly dependent on the data set used to train the algorithm. Therefore, one should consider the target population for which the application is being developed and train on a similar data set with the same pretest probability to have OSA.

Next to the Apnea-ECG data set, the NSRR [34] also provides PSG data sets,

of which the SHHS [108] has been used the most for the automatic detection of sleep apnea. This data set, however, was recorded in the general population, whereas sleep experts advise against OSA screening in the general population, but rather advise to screen groups with a high pretest probability in order to ensure a high post-test probability [30].

The availability of large public data sets, ideally including wearable data recorded simultaneously with the gold standard PSG, is even more important for algorithms based on deep learning. These algorithms are completely data-driven and their performance tends to improve if more data is available.

Another important factor when comparing different algorithms is the testing methodology. In order to obtain a fair generalization performance, the test set should be patient independent. However, studies often do not mention specifically whether subject independence of the test set was imposed [97]. Especially for cardiovascular signals the classification performance can increase significantly if part of the data from a test set patient was included in the training set, since cardiovascular features can be highly patient dependent.

3.9 Overview of OSA screening devices

Several portable devices have been proposed for the screening of OSA within research projects, and as commercial devices. These wearable solutions range from out-of-center polygraphy devices, which still include most of the PSG signals except for the EEG, to wearables which only measure one or two modalities. For this range of devices, a trade-off should be made between recording more modalities, to increase the screening performance, and the comfort of the patient.

Most producers of in-hospital PSG systems offer an out-of-center polygraphy alternative, Respironics, for example, has the Alice NightOne [112], whereas Medatec offers the Dreamscan 2 [84]. Although these devices provide good quality SpO_2 and respiratory signals, they are still quite uncomfortable for the patient.

Alternative single-sensor SpO_2 systems have thus been proposed, such as the Nonin WristOx which consist of a finger pulse oximeter connected to a wrist-worn device [98]. Within such device, additional modalities could be measured, such as the PAT and actigraphy included in the WatchPAT [65]. Moreover, several research projects have proposed the use of a finger pulse oximetry which sends the recorded data to a smartphone via Bluetooth [85]. Beddr even

proposes an oximetry device placed on the forehead with their SleepTuner [20]. Popular producers of smartwatches, such as Fitbit, Samsung and Apple include a pulse oximeter to measure PPG signals. However, none of them have been able to implement reliable SpO_2 monitoring in their watches. Studies have shown that the SpO_2 could only be reliably recorded on the finger and the forehead [80], technological advances in sensor design will thus be needed in order to implement this measure in smartwatches.

Pressure sensor devices to be embedded in the mattress are the most popular unobtrusive respiratory systems. The Beddit and Emfit sensors are two commercial examples of this type of sensor [6, 47].

A more extensive review on research and commercial OSA screening devices can be found in [85].

3.10 Future of sleep apnea screening: beyond the AHI?

Most of the studies developing automated screening algorithms for OSA validate their methods based on the AHI, which is used to define OSA severity in clinical practice. The AHI, however, only takes into account the number of apneas, but not their severity. This might explain why the AHI does not correlate well with the sleepiness of patients nor with cardiovascular outcome [56, 70, 121]. Moreover, when an AHI cut-off of 15 events per hour of sleep is considered, prevalences of up to 50 % can be observed in middle-aged male populations [14]. Therefore, researchers are searching for alternatives for the AHI in order to better phenotype OSA subjects and prioritize them for treatment.

In [44], Eckert reviews phenotypic approaches for OSA, with a focus on the PALM score. This score describes 4 traits causing OSA: an impaired upper airway anatomy, a low respiratory arousal threshold, an unstable ventilatory control system and ineffective upper airway dilator muscles. PALM is an acronym for the measured parameters for each of the 4 studied traits: **P**crit, **a**rousal threshold, **l**oop gain and **m**uscle responsiveness. According to the most dominant trait, different treatments could be proposed. Measures of each of the causes were defined based on gold standard PSG and CPAP drops, but should still be tested on wearable measurements.

In [70, 126, 138], on the other hand, SpO_2 severity parameters are proposed which correlate better than the AHI with the cardiovascular risk of OSA subjects.

This area of research is promising to enhance phenotyping of OSA subjects

and in the future, these developed parameters could also be extracted from wearables in a home environment.

Part II

Algorithms

Chapter 4

Classification

Before going into the applications for OSA detection and phenotyping, the next two chapters will focus on the machine learning techniques used and developed during this PhD. This chapter discusses the different supervised classification algorithms and their performance evaluation.

One of the main algorithmic contributions of this PhD will be presented, namely the extension to multilevel ICS, which was published in: ***Deviaene, M.****, Borzée, P., Van Gilst, M., Van Dijk, J., Overeem, S., Buyse, B., Testelmans, D., Van Huffel, S., and Varon, C. Multilevel interval coded scoring to assess the cardiovascular status of sleep apnea patients using oxygen saturation markers. IEEE Transactions on Biomedical Engineering (2020).*

4.1 Introduction on classification methods

Machine learning is the field of study in which algorithms are developed which can learn to detect specific patterns or predict certain events. Machine learning applications can be divided into two main categories: supervised learning and unsupervised learning. In supervised learning prior knowledge is available for the training set. The expected outcome of the model is known and the model can be optimized based on the training set performance. Whereas, for unsupervised learning the training set is unlabeled. The advantage of unsupervised methods is the fact that there is no need to annotate the training data, it is thus easier to gather large data sets for training. However, when prior knowledge is available, supervised methods tend to obtain better performances. In this

thesis, only supervised methods were applied. As for the development of apnea detection algorithms, the labels of the training data were available since PSGs are manually annotated during standard clinical practice. For the phenotyping of OSA patients, according to their cardiovascular status, data sets were created in which cardiovascular labels are available, therefore supervised methods were applied. However, this assessment of the cardiovascular status of OSA subjects is generally not performed, so for other data sets the use of unsupervised methods might be an interesting research opportunity.

Supervised learning techniques can be further divided into regression and classification problems. In regression problems, such as time series predictions, the predicted outcome is a value from a continuous interval. Whereas in classification problems, the predicted class label is chosen from a limited number of possible classes. Only supervised classification was used in this thesis. For the apnea detection task, time windows were classified as apneic or not. Whereas for the OSA phenotyping task, subjects were divided into three cardiovascular comorbidity classes, which had to be predicted.

In the next section, different supervised classification methods from the literature will be presented, followed by sections on the extension to multilevel ICS models and a comparison between the different models. The last section will discuss the performance evaluation of classifiers.

4.2 Feature-based classifiers

4.2.1 k-nearest neighbors

The kNN algorithm will compare a new data point to all points from the training set. The k training points with the minimal distance to the new data point are considered the nearest neighbors, and the new data point receives the class label that is most common in these nearest neighbors [31]. Different distance measures can be applied, but the Euclidian distance is most commonly used. The only hyperparameter which should be chosen within the kNN model is the number of nearest neighbors k. When k would be taken as 1, the data point would be assigned the class label of the closest training sample. But this point might be an outlier, not representative for its class, therefore, to make the classifier more robust, a larger value for k is recommended [110]. But k should not be taken too large, otherwise the classifier might be smoothed too much, and might not capture the full complexity of the data set.

4.2.2 Linear discriminant analysis

Classification problems can be seen as finding the optimal way to divide the input space of features into regions labeled according to the class labels. The decision boundary between two classes represents the hyperplane in which the probability for a sample to belong to either one of the classes is equal. The estimation of these probabilities will be used to define the optimal decision boundary. The class probability is dependent on the prior probability of samples belonging to each class, π_k and the class density $f_{k(x)}$. When using an LDA classifier, it is assumed that the density of each class can be modeled as Gaussian densities with equal covariance matrices. Thanks to these assumptions, the LDA decision boundaries simplify to linear combinations of the input features [59]. In the training phase, the LDA classifier will fit Gaussian distributions to the training data for each class in order to estimate the decision boundaries.

Even though the data of real-life applications does not comply with the assumptions made to optimize the LDA decision boundaries, often good classification performances are obtained [59].

4.2.3 Random forest

A random forest classifier is made up of an ensemble of decision trees. Decision trees are rule-based algorithms, which will consecutively split up the feature space in smaller regions, until each final region only contains training samples from a single class [59]. This simple decision tree model, however, tends to overfit to the training data leading to a poor generalization on new test samples. Pruning and early stopping have been implemented to avoid this [74], but an even more powerful solution is to combine different decision trees trained on sub-parts of the training set. All trees will cast a class prediction, and majority voting is used to assign the final class label.

As such, a random forest classifier will train different decision trees of which the training hyperparameters are independent identically distributed random vectors in order to decrease the correlation between the trees [21]. This is obtained by training each tree on a random subset of data points (bagging) and by performing a random sampling of features at each node of the tree [21]. The samples which are not used for training per tree are called out-of-bag samples (OOB) and will be used to assess the validation performance of the classifier.

4.2.4 (Least-squares) Support vector machines

This section explains the use of SVMs for classification problems and is mainly based on [123], to which the reader is referred to for further reading on this topic.

Linear SVM

Support vector machine is a binary classification model in which the separating hyperplane is optimized by maximizing the margin between the two classes for linearly separable data. The margin is defined as the distance between the nearest two points of different classes. The input data set will be rescaled such that these nearest points satisfy $|\mathbf{w}^T\mathbf{x}_i + b| = 1$. As such, the margin to be maximized is equal to $2/||\mathbf{w}||_2$. This corresponds to the minimization of $||\mathbf{w}||_2$. If a linearly separable training set is available with N input data vectors $\mathbf{x}_i \in \mathbb{R}^{N_d}$ and class labels $y_i \in \{-1, +1\}$, a classifier can be built that perfectly separates the two classes by optimizing:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, ..., N. \end{aligned} \tag{4.1}$$

Real-life data, however, is almost never linearly separable, therefore slack variables ξ_i are introduced in order to allow the misclassification of a limited number of samples. The regularization parameter c will balance maximizing the margin and limiting the number of misclassifications, which leads to the following optimization function:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + c\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & \begin{cases} y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad i = 1, ..., N. \end{aligned} \tag{4.2}$$

Use of kernels for nonlinear SVM

In order to extend the linear SVM classification to nonlinear solutions, the input data is mapped to a higher dimensional feature space using a nonlinear mapping $\varphi(\mathbf{x})$. A linear separating hyperplane will then be constructed in this new feature space following Equation 4.2, provided $\mathbf{x}_i$ is replaced by $\varphi(\mathbf{x}_i)$.

This new optimization function is called the primal problem and is solved by formulating the Lagrangian and computing the saddle point, which leads to $\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \varphi(\mathbf{x}_i)$, with α_i the Lagrange multipliers.

Since the calculation of this nonlinear mapping is computationally expensive and even impossible in some cases, the kernel trick is applied, which enables this problem to be solved without explicitly calculating $\varphi(\mathbf{x})$. Each positive definite function can be chosen as kernel $K(\mathbf{x}, \mathbf{z})$ with $\mathbf{x}, \mathbf{z} \in \mathbb{R}^{N_d}$, so that $K(\mathbf{x}, \mathbf{z}) = \varphi(\mathbf{x})^T \varphi(\mathbf{z})$. Popular choices are linear, polynomial, and radial basis function (RBF) kernels. When the kernel trick is used, the optimization function in the dual space becomes:

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2} \sum_{i,j=1}^{N} y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \alpha_i \alpha_j + \sum_{i=1}^{N} \alpha_i \\ \text{s.t.} \quad & \begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq c \end{cases} \quad i = 1, ..., N. \end{aligned} \tag{4.3}$$

As a result, the nonlinear SVM classifier is achieved:

$$y(\mathbf{x}) = \text{sign}\left[\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \right]. \tag{4.4}$$

The solution will be sparse, as many Lagrange multipliers will be equal to zero. The data points with non-zero α's are called support vectors, and they define the decision boundary [123].

Least-squares SVM (LS-SVM)

A further simplification of the SVM model was proposed by Suykens et al. [123], in which the cost function is defined as a least squares problem and the inequality constraints are replaced by equality constraints. The LS-SVM primal classification problem is given by:

$$\begin{aligned} \min_{\mathbf{w},b,e} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \sum_{i=1}^{N} e_i^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 - e_i \quad i = 1, ..., N, \end{aligned} \tag{4.5}$$

in which γ is a regularization constant. This leads to the following linear system:

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^T \\ \hline \mathbf{y} & \boldsymbol{\Omega} + \mathbf{I}/\gamma \end{array} \right] \left[\begin{array}{c} b \\ \hline \boldsymbol{\alpha} \end{array} \right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1_N} \end{array} \right], \tag{4.6}$$

where $\mathbf{y}, \boldsymbol{\alpha}$ and $\mathbf{1_N}$ are column vectors of size 1×N including, respectively, the class labels, the Lagrange multipliers, and ones. Additionally, $\boldsymbol{\Omega}$ is equal to $y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$. This set of linear equations is computationally less expensive to solve, than the quadratic programming problem given for the standard SVM in equation 4.3. The LS-SVM has as a disadvantage that it lacks the sparseness present in the standard SVM solution. In other words, no α's will be equal to 0 and all training samples will contribute to the construction of the separating hyperplane. When large training sets are available it is, therefore, not recommended and often unfeasible, to train on the full training set. Fixed-size LS-SVM (FS LS-SVM) has been introduced to overcome this limitation. FS LS-SVM, makes an approximation of the LS-SVM model according to the Nyström method, assuming a fixed size of M support vectors. These M support vectors will be selected from the training set according to the entropy criterion [123]. In the original implementation, the initial M support vectors are chosen at random. In order to speed up the convergence of this sample selection, an alternative initialization procedure was proposed in [129], which was used in this work. In this case, the M support vectors are chosen as centroids of M clusters, computed using k-means.

4.2.5 Interval coded scoring

Interval coded scoring models are used to automatically develop interpretable risk score models for clinical decision support systems [18]. This paragraph will give a general overview of the ICS algorithm, the different steps will be discussed in more detail in the following paragraphs. The ICS algorithm starts by dividing each variable into a number of discrete intervals. Next, an optimization algorithm is applied to assign a score value to each of these intervals and iterative reweighting is applied to further simplify the obtained model. The risk score for a sample can then be computed by summing the score values of the final model over the different variables. The empirical risk linked to this score is also provided by the ICS model.

Before applying the core ICS algorithm, the input data matrix $\mathbf{X} \in \mathbb{R}^{N \times N_d}$, with N the number of training samples and N_d the number of features, is transformed into the binary matrix $\mathbf{Z} \in \mathbb{B}^{N \times N_s}$. The matrix $\mathbf{Z}$ is constructed by dividing the feature values into a number of discrete intervals. The interval thresholds are calculated such that each interval contains the same number of training samples, with a minimum of 5 samples. The default maximum number of intervals used is 10. For each of the obtained intervals, a binary vector $\mathbf{z}$ is defined, which is 0 for all samples, except those which feature values lie within this interval. N_s corresponds to the total number of intervals over all variables.

Next, ICS optimization is applied to assign a score to each interval of every feature. The ICS toolbox, publicly available at [49], provides two optimization techniques, one based on linear programming and another one using elastic nets. In this PhD thesis only the linear programming ICS model was considered, and the focus of this section will thus be on this implementation. This ICS optimization is based on the soft-margin SVM presented in Equation 4.2, but the maximal margin objective of the SVM was replaced by total variation minimization [18] using the objective term $||\mathbf{Dw}||_1$. $\mathbf{D}$ is a binary matrix which encodes the differences between coefficients of adjacent intervals of the same parameter. As such, the number of weight changes is limited between adjacent intervals, leading to a simpler scoring model which focusses on general trends rather than dataset-specific details. This leads to the following linear programming problem:

$$\begin{aligned} &\min_{\mathbf{w},b,\boldsymbol{\varepsilon}} \quad ||\mathbf{Dw}||_1 + \gamma \boldsymbol{\varepsilon}^T \mathbf{1} \\ &\text{s.t.} \quad \begin{cases} y_i(\mathbf{z}_i \mathbf{w} + b) \geq 1 - \varepsilon_i \\ \varepsilon_i \geq 0 \end{cases} \quad i = 1, ..., N. \end{aligned} \tag{4.7}$$

In this formulation, $\mathbf{w}$ and b represent the coefficients and the bias of the model, respectively. $\boldsymbol{\varepsilon}$ is the vector of slack variables used for regularization controlled by the regularization constant γ. This problem can be restated by introducing a new vector $\mathbf{f} = |\mathbf{Dw}|$:

$$\begin{aligned} &\min_{\mathbf{w},b,\mathbf{f},\boldsymbol{\varepsilon}} \quad \mathbf{f}^T \mathbf{1} + \gamma \boldsymbol{\varepsilon}^T \mathbf{1} \\ &s.t. \quad \begin{cases} \mathbf{Dw} - \mathbf{f} \leq \mathbf{0} \\ -\mathbf{Dw} - \mathbf{f} \leq \mathbf{0} \\ y_i(\mathbf{z}_i \mathbf{w} + b) \geq 1 - \varepsilon_i \\ \varepsilon_i \geq 0. \end{cases} \quad i = 1, ..., N. \end{aligned} \tag{4.8}$$

Next, iterative reweighting is applied to further simplify the model. This is done by penalizing small weight differences between adjacent intervals in order to force them to be smoothed out in the next iteration [18]. Neighboring intervals with the same weight will be merged together, and irrelevant features might also be removed if all discrete intervals end up with the same weight. The loop of ICS model optimization and iterative reweighting is repeated until convergence is reached with a maximum of 10 iterations. The balance between model simplification and performance can be steered by the user. A relative AUC cut-off value can be defined as hyperparameter, which will constrain the reweighting step. Model simplifications resulting in a relative decrease in AUC smaller than this threshold will be allowed.

In the end, the weight vector $\mathbf{w}$ is rescaled to obtain integer score points $\mathbf{s}$. The total risk score of a sample can then be obtained by summing the scores of the corresponding intervals for each relevant parameter. Logistic regression is then applied to link the risk scores to the empirical risk of having the studied condition.

The ICS toolbox [49] also provides the possibility to use the interactions between two parameters as input. A matrix with discrete intervals for both parameters is then created. Additionally, an elastic net parameter pre-selection method is available, which can remove irrelevant parameters and interactions before computing the ICS model to reduce computation time [18]. Like the SVM classification models, ICS models were only designed for binary classification. In Section 4.3.1 an extension to multilevel classification is presented which was developed within this PhD.

4.3 Multilevel ICS

4.3.1 Extension to multilevel ICS

The standard ICS implementation described in the previous section only supports 2-class classification, with y_i consisting of binary labels. In the apnea phenotyping example, however, three cardiovascular severity categories were defined, Ct, cC and cE. When a hierarchical ordering is present between the different severity categories, and this hierarchy is consistent throughout the used features, a 3-class classification is possible by optimizing only two separating hyperplanes, as can be seen from the example in Figure 4.1. The first hyperplane will, in this study, separate the Ct group from both the cC and cE groups, whereas the second hyperplane will separate the cE group from the Ct and cC groups. The main assumption on the data for this model to work is the fact that the hierarchy of the categories should be kept throughout the feature space, in this case, this means that the cC group feature values should always lie in between the Ct and cE group values.

Additionally, the proposed mICS approach aims to output one final model which can capture the different levels of severity in the data set using the same mICS scores. Therefore, the coefficients $\mathbf{w}$ will be kept the same for both hyperplanes, whereas the bias b will determine the difference between the two hyperplanes. Also, two regularization constants γ will be optimized.

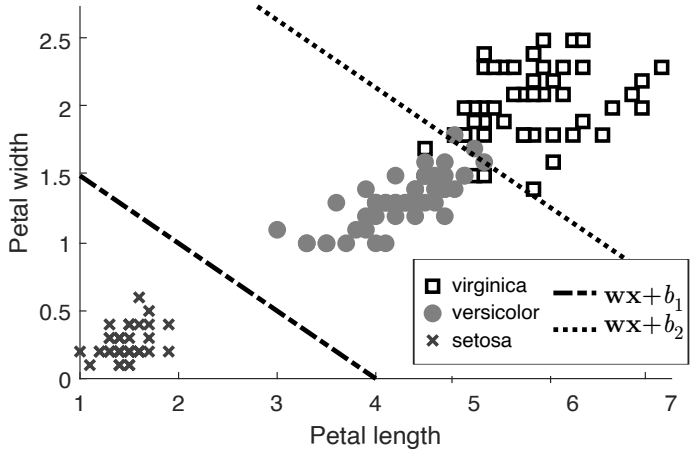


Figure 4.1: Example of finding the optimal hyperplanes for a multilevel problem, data extracted from the Fisher's Iris data set.

These adaptations will result in the following linear programming problem:

$$\min_{\mathbf{w},b_1,b_2,\mathbf{f},\boldsymbol{\varepsilon_1},\boldsymbol{\varepsilon_2}} \quad \mathbf{f}^T\mathbf{1} + \gamma_1\boldsymbol{\varepsilon_1}^T\mathbf{1} + \gamma_2\boldsymbol{\varepsilon_2}^T\mathbf{1}$$

$$s.t. \quad \begin{cases} \mathbf{Dw} - \mathbf{f} \leq \mathbf{0} \\ -\mathbf{Dw} - \mathbf{f} \leq \mathbf{0} \\ y_{i,1}(\mathbf{z}_i\mathbf{w} + b_1) \geq 1 - \varepsilon_{i,1} \\ y_{i,2}(\mathbf{z}_i\mathbf{w} + b_2) \geq 1 - \varepsilon_{i,2} \\ \varepsilon_{i,2} \geq 0 \\ \varepsilon_{i,2} \geq 0 \end{cases} \quad i = 1, ..., N. \tag{4.9}$$

4.3.2 mICS model with monotonicity constraints

When only a small set of training samples is available, the ICS optimizer might end up with a model in which some variables have unexpected dips or peaks. This is due to an overfitting of the training data, leading to these dataset-specific details being included in the model. However, in order to improve the interpretability of the model, the scoring model should mainly focus on general trends. These trends are often expected to be monotonical per variable. Overfitting of the ICS model can thus be avoided by adding an extra monotonicity constraint to the weights. In order to ensure monotonicity of the mICS interval scores per parameter, the constraints of the mICS weight optimization were adapted. The first two constraints of Equation 4.9 will be

adjusted to ensure that the differences between coefficients always have the right sign. These constraints then become:

$$\begin{cases} \mathbf{DwS} \leq \mathbf{0} \\ -\mathbf{DwS} - \mathbf{f} = \mathbf{0} \end{cases} \tag{4.10}$$

The new matrix $\mathbf{S}$ contains -1's for parameters that have a positive correlation with an increasing risk, and +1's if the parameters are negatively correlated with the risk.

4.4 Comparison of methods

When comparing the presented classification methods in terms of classification accuracy, the random forest and SVM models are generally considered the most powerful ones [74]. But this does not mean that they always outperform the other methods on each application. In some cases LDA, for example, can perform equally good or even better [88]. RF and SVM classifiers have the ability to construct more complex decision boundaries, but this also entails a greater risk to overfit the model to the training set, leading to a poor generalization to new samples. Moreover, the final classification model will be difficult to interpret and is often considered as a black-box. The ICS model, on the other hand, has been developed to be as interpretable as possible. This increased interpretability can sometimes cause a slight drop in accuracy, but especially in biomedical applications interpretability of methods is very important. Therefore, deep learning methods were not considered in this PhD thesis, as they are even harder to interpret than RF and SVM classifiers. Moreover, these deep learning methods require large data sets. The number of available recordings used in Chapter 10 for OSA screening using unobtrusive measurements, and in Chapters 11 and 12 for OSA phenotyping, is not enough to train deep learning methods.

Another important parameter is how well the classifier deals with irrelevant features and outliers. kNN is unfortunately not well at dealing with both, so implementing good data preprocessing steps is crucial when selecting this algorithm [74]. However, kNN can easily be used for incremental learning, in which a pre-trained classification model gets adapted using new training samples [74]. It also has a fast training time, and includes only one hyperparameter to be tuned, i.e. the number of nearest neighbors. Whereas, SVM, random forest and ICS models all have more hyperparameters to be tuned and are computationally more complex to be trained, especially when the number of input features increases.

All classifiers have their advantages and disadvantage, so the classification model should be chosen according to the requirements of the problem at hand. If the only goal is to maximize the classification accuracy, it can be recommended to test different models.

4.5 Performance evaluation

In order to assess the performance of a classifier, the predicted class labels will be compared to the true class labels. From this comparison, different performance measures can be computed. The use of a confusion matrix, receiver operating characteristic (ROC) curve and kappa score will be discussed in this section. Moreover, in order to assure the generalization of the classifier to unseen data samples, the data set should be split into training, validation and test sets. The procedures used in this work for splitting the data will be discussed in the last subsection.

4.5.1 Confusion matrix

In a confusion matrix, each cell represents the number of samples predicted to belong to a certain class for each possible true class label. The confusion matrix for a two-class problem is given in Table 4.1. In this matrix four groups are formed: true positive (TP), false positive (FP), false negative (FN) and true negative (TN) samples. Using the number of samples in each of these groups, several performance metrics can be computed:

- Sensitivity (Se): $\frac{TP}{TP+FN}$
- Specificity (Sp): $\frac{TN}{TN+FP}$

Table 4.1: Confusion matrix for a two class problem. Correctly classified groups are indicated in white, while misclassified groups are colored in gray.

		True label		
		Positive	Negative	
Predicted label	Positive	TP	FP	→ PPV
	Negative	FN	TN	→ NPV
		↓	↓	
		Se	Sp, FPR	

- Accuracy (Acc): $\frac{TP+TN}{TP+FN+TN+FP}$
- Positive predictive value (PPV): $\frac{TP}{TP+FP}$
- Negative predictive value (NPV): $\frac{TN}{TN+FN}$
- False positive rate (FPR): 1 - Sp = $\frac{FP}{FP+TN}$

4.5.2 Receiver operating characteristic curve

The receiver operating curve is shown in Figure 4.2. It plots the sensitivity with respect to the false positive rate for different operating points of the classifier. These operating points are obtained by varying the decision threshold on the posterior probability output of the classifier. An optimal operating point on the AUC can be chosen to obtain a trade-off between the number of false positives and false negatives. Moreover, the area under the ROC is used as a performance measure. The larger this AUC, the better the classifier. A classifier with an AUC of 0.5 represents a random classifier.

4.5.3 kappa score

In unbalanced or multi-class problems, the Cohen's kappa value is a more appropriate performance measure [13]. The kappa value was designed as

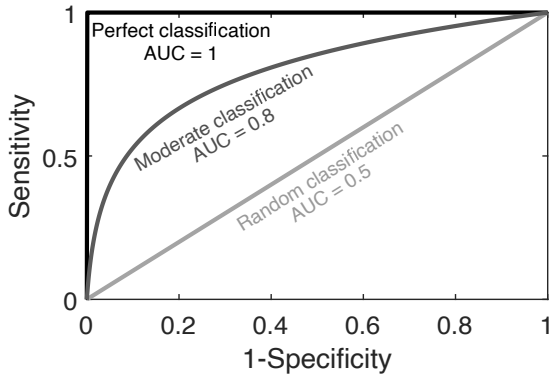


Figure 4.2: ROC curve for different cases. Area under ROC curve (AUC) serves as a performance parameter.

a measure of inter-rater agreement and gives an estimate of how good the agreement is between the labels of two raters, or in this case between the gold standard class labels and the predicted classes, with respect to the degree of agreement that can be expected due to chance. This value is computed as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{4.11}$$

with p_o the observed agreement and p_e the expected agreement. p_o is equal to the accuracy of the classifier, whereas p_e is computed as the sum, over the different classes, of the multiplication of the ground truth class proportion and the predicted class proportion. By comparing with the chance level, the method accounts for unbalanced data and this also enables the comparison of performances between data sets with different prior class distributions.

4.5.4 Training, validation and test set

The performance metrics discussed in the previous subsections should be tested on unseen data in order to assure that the classifier will generalize well on new data samples. Therefore, data sets are usually split into a training set, used to train the classifier, and an independent test set which cannot be used in any way to optimize the classifier. Additionally, a third validation set is often introduced in order to avoid overfitting to the training set. When a classifier overfits, an almost perfect classification performance is obtained on the training set, while a poor performance is achieved on the test set. The hyperparameters of the classifier, e.g. k the number of neighbors in kNN, or the kernel type and regularization parameter c in SVM, will then be chosen such that the performance on this validation set is optimized.

The use of a separate validation set, has as a disadvantage that the obtained optimal hyperparameter values might be dependent on the choice of samples included in the validation set. Moreover, by dividing the data set in three, the number of samples available for training is further reduced. Therefore, many classification algorithms make use of n-fold cross-validation instead. In this case, the training set is divided into n groups, and the classifier is retrained n times. Each iteration, the classifier is trained based on samples from n-1 groups and tested on the group that was left out. By averaging the n obtained test set performances, a cross-validated performance can be obtained. Typically, 10-fold cross-validation is used. But when only a small data set is available, leave-one-out cross-validation can be preferred, in which the number of folds is equal to the number of training samples.

In biomedical signal processing problems, often multiple data samples are available per subject. These samples might contain characteristics specific to

the studied subject. Including samples of a subject in both the training and test set can thus lead to a significant increase in performance. Therefore, to ensure the generalization of the classifier on new subjects, it is recommended to split the data into training, testing and validation sets per patient instead of sample-wise. When the number of subjects is limited, leave-one-patient-out cross-validation (LOPO CV) is also possible.

4.6 Conclusion

This chapter introduced the different supervised classification algorithms that were used in this research, together with the different performance metrics used to evaluate them. More importantly, one of the contributions of this PhD was presented, namely an extension to the ICS algorithm. This extension to multilevel ICS is not only useful for this research. In fact, it is applicable to any multiclass problem, where a hierarchy can be imposed between the classes. In addition, this extension was made publicly available in [49].

These methods will be used in Part III to develop algorithms for the automatic detection of apneic events and in Part IV to assess the cardiovascular status of OSA subjects. But first, the next chapter will discuss feature selection methods that choose the most relevant features to be given as input to the classifier.

Chapter 5

Feature selection

The previous chapter described different supervised classification models, but before these classifiers can be applied, the most relevant feature set should be identified using feature selection methods. This will be the topic of this chapter. First a short introduction will be given, after which several feature selection methods from the literature will be discussed. Then, a novel feature selection method based on random forest will be presented, this corresponds to another algorithmic contribution of this PhD. To conclude, the different feature selection methods will be compared.

The novel RF-based feature selection method was published in: ***Deviaene, M.****, Testelmans, D., Borzée, P., Buyse, B., Van Huffel, S., and Varon, C. Feature selection algorithm based on random forest applied to sleep apnea detection. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019), IEEE, pp. 2580–2583.*

5.1 Introduction to feature selection

One of the most important problems in machine learning applications is to decide which features to use to build a model. Often, a large number of relevant features can be defined which might have an added value to the model. Some of these features could, however, be less informative than expected or highly correlated, leading to redundant features. Therefore, feature selection methods are used to select an optimal feature set which contains the most relevant features, while avoiding redundancy. This can avoid overfitting to the training

set and can lead to an improved performance. Moreover, a reduced feature set leads to a faster feature extraction and model optimization and will improve the interpretability of the model [58]. This is very important when developing clinical screening algorithms for wearable devices, since the computational cost should be as low as possible and clinicians are interested in the underlying physiological mechanisms of the machine learning model.

Multiple studies have focused on the development of feature selection methods, resulting in numerous algorithms. Generally speaking, feature selection methods can be grouped in three classes: filter methods, wrappers, and embedded methods [125]. The filter methods evaluate the features independent of the used classifier. In this chapter, the F-score will be presented as an example of a univariate filter, followed by the description of the minimal redundancy maximal relevance algorithm, which is a multivariate filter. The next subsection will discuss wrapper methods, which construct the feature set in a sequential approach based on the resulting classification performance. Afterwards, the novel random forest based feature selection method, proposed in this thesis, will be introduced, which is an embedded method. The chapter will conclude with a comparison of the different methods.

5.2 Feature selection methods from the literature

5.2.1 F-score

The F-score is a univariate filter, which ranks the features independently of each other. The F-score uses the ratio between the intra-group variability and the inter-group variability, based on the Fisher criterion, which is also the basis for LDA classification (see Section 4.2.2). The score is computed for the ith feature as:

$$F(\mathbf{x}^i) = \frac{(\mu_i^+ - \mu_i)^2 + (\mu_i^- - \mu_i)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}(x_{k,i}^+ - \mu_i^+)^2 + \frac{1}{n^--1}\sum_{k=1}^{n^-}(x_{k,i}^- - \mu_i^-)^2}, \tag{5.1}$$

in which μ_i, μ_i^+ and μ_i^- represent, respectively, the average values of the ith feature for the whole set, the positive class, and the negative class [129]. n^+ and n^- represent the number of, respectively, positive and negative samples. The larger the F-score, the larger the discriminative power of the feature.

5.2.2 Minimal redundancy maximal relevance criterion

The mRMR criterion [103] is a multivariate filter measure which combines selecting the most relevant features and reducing the redundancy among those features. It will build the optimal feature set by sequentially adding one feature at a time to the optimal feature subset $\mathbf{S}$. The dependency between two variables $\mathbf{x}$ and $\mathbf{y}$ will be characterized by their mutual information $I(\mathbf{x};\mathbf{y})$ which is defined as:

$$I(\mathbf{x};\mathbf{y}) = \int\int p(\mathbf{x},\mathbf{y}) log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x} d\mathbf{y}, \tag{5.2}$$

in which $p(\mathbf{x})$ and $p(\mathbf{y})$ represent the probabilistic density functions of, respectively, $\mathbf{x}$ and $\mathbf{y}$, and $p(\mathbf{x},\mathbf{y})$ is the joint probability density function.

The feature with the maximal relevance to the target class $\mathbf{c}$ can be found by selecting the feature $\mathbf{x}_j$ with the largest mutual information $I(\mathbf{x}_j;\mathbf{c})$. Minimizing the redundancy within the feature set can be interpreted as minimizing the dependency, and thus the mutual information, between feature $\mathbf{x}_j$ and all features $\mathbf{x}_i$ that already belong to the feature subset $\mathbf{S}$: $\sum_{\mathbf{x}_i \in \mathbf{S}} I(\mathbf{x}_j;\mathbf{x}_i)$. The incremental mRMR algorithm will combine those two optimization functions. When the optimal feature subset $\mathbf{S}$ contains $m-1$ features, the mth feature selected from feature set $\mathbf{X}$ will optimize the following condition:

$$\max_{\mathbf{x}_j \in \mathbf{X} \setminus \mathbf{S}_{m-1}} \left[I(\mathbf{x}_j;\mathbf{c}) - \frac{1}{m-1} \sum_{\mathbf{x}_i \in \mathbf{S}_{m-1}} I(\mathbf{x}_j;\mathbf{x}_i) \right]. \tag{5.3}$$

In [103], it was proven that this mRMR criterion is equivalent to the maximal dependency criterion if only one feature is selected at a time. The method, however, does not specify the optimal number of features. Since the method incrementally selects features, once a feature is selected it cannot be removed again. Therefore, the final feature set can still contain redundant features. It is advised to use the mRMR method to find a small set of candidate features and further reduce this set with a backward wrapper, which will be discussed in the next subsection.

5.2.3 Wrapper methods

Wrapper methods make use of the obtained classification performance on a specific classification model to find the optimal subset of features. Backward and forward wrappers exist. The backward wrapper starts from the full feature set containing N features and will sequentially remove one feature at each iteration.

Whereas, a forward wrapper starts with an empty feature subset, in which at each iteration, a feature is added.

In the case of a forward wrapper, in the $n+1^{th}$ iteration, the optimal feature subset already consists of n features, and N-n features remain in the data set. To decide which feature should be added next, the classifier will be retrained N-n times for each possible feature subset in which one additional feature is added to the optimal set. The classification performances of these classifiers will be compared, and the feature whose addition resulted in the biggest increase in performance will be selected. This process is continued until the classification performance does not increase any more. In the case of a backward wrapper, at each iteration, the feature whose removal resulted in the largest gain in performance will be removed.

The choice between a forward or a backward wrapper will depend on the number of features included in the data set. When the number of features is large, a backward wrapper might be computationally too expensive. Wrappers can be combined with any classification method, and any performance metric discussed in Chapter 4.

5.3 Random forest based feature selection

As proposed in [21], an embedded feature selection within a RF classifier can be performed by building a backward wrapper, in which for each iteration the feature with the lowest predictor importance (PI) is removed. This PI is calculated for each feature as the increase in OOB misclassification error when the values of the feature under investigation were randomly permuted. When the feature has a high relevance, the misclassification error will increase, which results in a high PI. If the feature only contains redundant information, also captured by other features, the misclassification error will not change much and the PI will be low. This is computationally more efficient than applying a normal backward wrapper, since the model should only be retrained once per iteration, instead of for all considered feature sets. A graphical overview of this random forest based feature selection method is given in Figure 5.1.

Further improvements to this RF feature selection method were proposed in this PhD. These were presented in [40], and will be discussed in this section.

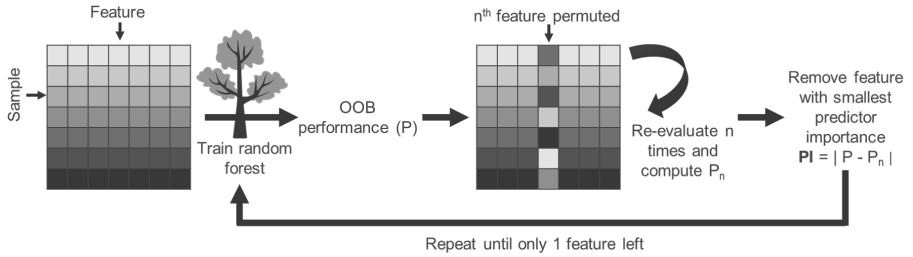


Figure 5.1: Schematic overview of the random forest-based feature selection.

5.3.1 Patient independent validation

In many clinical applications, feature values might be patient dependent with large inter-patient variabilities. These patient dependencies should be avoided, and the validation performance should take into account the generalizability of the classifier to new, unseen patients. To account for this, the selection of the OOB validation samples in the RF was adapted. Instead of selecting the OOB samples randomly, one third of the patients were randomly selected for each tree, and all samples of these patients were used as OOB samples. As such, the OOB prediction for each sample is only based on the trees that were trained without using any information of that patient, thereby avoiding overfitting.

5.3.2 Cohen's kappa value as performance estimate

In the standard RF feature selection method, the misclassification error is used to assess the performance. In unbalanced or multi-class problems the Cohen's kappa value is, as discussed in Section 4.5.3, a more appropriate performance measure. In the novel RF-based feature selector, it is thus proposed to use the Cohen's kappa value as performance measure.

5.3.3 Inclusion of feature correlation

It is expected that the standard RF feature selection should be able to remove correlated features. Since, in the case of correlated features, the PI should not alter much, as the RF will obtain the information contained in the feature under investigation from the correlated feature instead. However, studies have shown that the PI tends to be overestimated for correlated features [122]. This might be due to the random sampling of features at each node, which could cause the correlated feature to be excluded. To avoid the selection of redundant features,

this thesis proposes to multiply the predictor importance with one minus the maximum absolute Spearman correlation coefficient (C_{max}) between the feature under investigation and the other remaining features as follows:

$$PI = \Delta\kappa \times (1 - C_{max}), \tag{5.4}$$

with $\Delta\kappa$ the change in κ value. The PI of features that are highly correlated with another feature will then be further decreased, this will lead to more feature pruning. Moreover, to speed up computations, highly correlated features are removed at the beginning of the algorithm. For feature pairs with a Spearman correlation coefficient higher than 0.9, only the feature with the highest F-score was retained.

5.3.4 Determination of the optimal number of features

In order to determine the optimal number of removed features, the mean OOB Cohen's kappa errors over all trees were compared. Additionally, this value was multiplied with one minus the average correlation coefficient in the corresponding feature set. The feature set with a maximal value on this kappa-correlation performance parameter was selected. The set with only one feature was excluded, except when this case would have the highest OOB kappa coefficient.

5.4 Qualitative comparison of feature selection methods

When comparing the different feature selection methods, it is clear that the F-score and mRMR filter methods have the advantage of being independently of the used classifier. Moreover, these methods are computationally less expensive than the wrapper or RF-based method. The F-score, however, only ranks the features based on their individual discriminative power and does not take into account the correlation between features. Whereas, the mRMR algorithm does not provide a clear determination of the optimal number of features, and can thus still retain redundant features. The wrapper methods, on the other hand, are computationally very expensive when the data set contains a large number of features. The proposed RF-based algorithm is computationally more efficient than a standard wrapper and will provide an optimal number of features. Moreover, the improved RF-based algorithm will result in a better pruned feature set, thus improving the interpretability of the model. This might, however, result in a slight decrease in performance. Finally, the improved RF-based model includes patient-independent validation, and since the kappa score

is used, it will be more appropriate for unbalanced and multi-class classification problems.

The added value of the novel RF-based feature selection will be demonstrated quantitatively in Section 6.5.1. This section shows a comparison between the novel RF-based feature selector and the use of the mRMR algorithm combined with a backwards wrapper. It is shown that using the novel RF-based algorithm, the feature set could be further pruned from six to three features with a minimal change in performance.

As each feature selection method has its advantages and disadvantages, it can be beneficial to apply several feature selection algorithms sequentially. Remaining redundant features in the mRMR algorithm, can for example be removed by adding a backward wrapper as a post-processing step [103]. Moreover, when a large feature set is available, the feature selection step can be accelerated by adding a pre-processing step which removes highly correlated features, as was proposed in the last paragraph of Section 5.3.3.

5.5 Conclusion

Within this chapter, feature selection methods from the literature, which will be used in the remainder of this thesis, were introduced. Additionally, improvements to the random-forest based feature selection method were presented. The improved RF-based feature selector is generally applicable and will be especially useful for unbalanced and/or multi-class problems. This novel method leads to a reduced feature set, enhancing the interpretability of the model.

Part III

Apnea Detection

Chapter 6

SpO_2-based sleep apnea detection

The next five chapters will present the development of automated OSA screening methods. This chapter will start with the use of the blood oxygen saturation signal for the detection of apneic events. In this study, an event-based approach was followed in which oxygen desaturations were detected and features were derived from these desaturations. Novel quasi-periodicity features were introduced, which can capture the recurrence of successive apneas. Using the most discriminative features, a classifier was built, which predicts if the desaturation was caused by an apneic event or not. Several feature selection methods proposed in Chapter 5 will be combined, and different classifiers from Chapter 4 will be compared. We hypothesize that when the most discriminative SpO_2 features from different categories, e.g. desaturation severity and quasi-periodicity, are combined, higher classification performances can be achieved.

Based on the classification of desaturations, the AHI could be estimated. Another point of novelty in this work is the fact that linear regression of the estimated AHI was used to inflate the AHI in order to account for respiratory events without oxygen desaturations and awake periods. The regressed AHI was then used to classify subjects in OSA severity categories. Publicly available data sets were used, in order to enable comparison with previous studies. The algorithm was also tested on UZ Leuven data sets to assess the generalization capabilities of the algorithm.

This algorithm was published in ***Deviaene, M.****, Testelmans, D., Buyse, B., Borzée, P., Van Huffel, S., and Varon, C. Automatic screening of sleep apnea*

patients based on the spo$_2$ signal. IEEE journal of biomedical and health informatics 23, 2 (2018), 607–617. Additional experiments will be included in Section 6.5.

The remainder of this chapter starts with an overview of the used data sets in Section 6.1, followed by the explanation of the complete algorithm in Section 6.2. Afterwards the results of the algorithm will be presented and analyzed in Section 6.3, Section 6.4 discusses the algorithm further, compares the obtained results with methods from the literature, and Section 6.5 introduces some additional experiments that were performed after the publication of this study. In the end, conclusions are presented in Section 6.6.

6.1 Data sets

The SpO_2-based algorithm was developed based on the SHHS data sets. All available baseline and follow-up recordings in this data set have more than 3 hours of usable SpO_2 signal, and were therefore included. The SHHS data set was split into 3 subsets for this study. 500 recordings were selected from the baseline PSGs for training the algorithm, these are called *SHHS1 train*. The remaining baseline recordings will be called *SHHS1 test* and the follow-up recordings *SHHS2*. The training samples were selected as the centroids of k clusters, computed using k-medoids clustering of the patient characteristics (age, BMI, AHI and smoking status), with k the number of training samples. This selection was performed on eight subsets, separating males and females and the four apnea severity classes. In each subset k/8 training samples were selected in order to assure an equal number of males and females, and an equal distribution over the apnea severity classes. As such, the underlying distribution of the patient population in SHHS1 is fully represented in the training set.

Moreover, the obtained algorithm was tested on the 8 recordings of the Physionet Apnea-ECG database which contain the SpO_2 signal, and the UZ Leuven matched data set. Additionally, the UZ Leuven OSA+ and Imec data set were used to perform extra experiments in Section 6.6. More information on these data sets can be found in Chapter 2.

6.2 Methods

The algorithm starts by preprocessing the SpO_2 signals. Afterwards, desaturations were detected. Class labels were then computed for these

desaturations, based on the presence of a preceding apnea or hypopnea that caused the desaturation. For each desaturation, 143 features were extracted, feature selection techniques were subsequently used to select the most discriminative features. Based on these features, different classifiers were trained to detect apneic events. Based on the number of detected events, an estimation of the AHI could be made in the end. The discussed steps are explained in detail in the following sections.

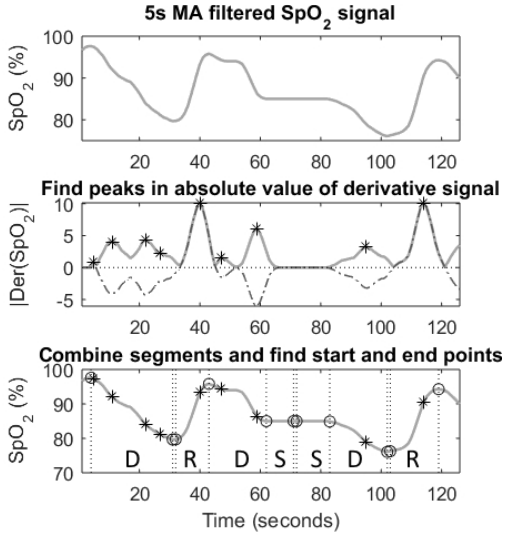
6.2.1 SpO$_2$ preprocessing

Before analyzing the signal, zero-level artifacts due to sensor disconnections were deleted by detecting desaturations in the SpO$_2$ that drop below 50 % and replacing them by linear interpolations. Afterwards, a moving average (MA) filter with a 3 seconds duration [72] was used to get rid of sharp changes and ripples due to oversampling in the UZ Leuven data sets, which were sampled at 500 Hz. This high sampling rate also caused the need to down sample the signals to 1 Hz, in order to equalize the sampling rates and speed up computations [72].

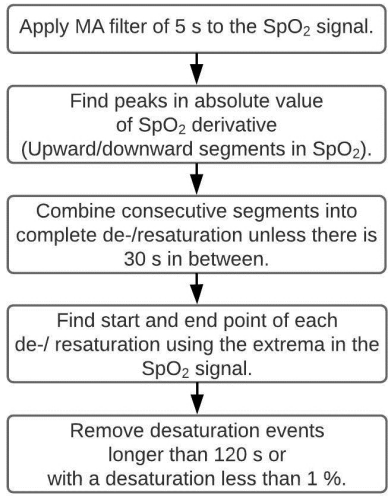
6.2.2 Desaturation event and baseline extraction

When an apneic event occurs, a drop in the blood oxygen concentration, or so-called desaturation (D), is often observed. Afterwards, possibly a stable period (S) of low SpO$_2$ occurs and then resaturation (R) normally brings the SpO$_2$ back to its baseline value, as can be seen in the lower graph of Figure 6.1a. The term *desaturation event* will be used to depict the complete event starting from the start of the desaturation until the end of the resaturation, as depicted in Figure 6.2. The goal of this study was to detect apneic events based on their desaturation event. The first step was thus to detect all desaturations in the SpO$_2$ signal, the different steps in this process are depicted in Figure 6.1.

In order to get a smoother signal and make the detection of desaturations easier, the first step was to apply an extra MA filter of 5 seconds to the preprocessed signal. This extra MA filter was solely used for the detection of desaturations, as the extra smoothing would affect the calculation of SpO$_2$ features. Next, derivative filtering was performed on the resulting signal. Peaks in the absolute value of the derivative were detected, representing upward and downward parts in the SpO$_2$ signal [95] (Figure 6.1a, middle). Consecutive downward segments were linked in order to create the complete desaturation, unless the detected peaks were more than 30 seconds apart [95]. In that case, it is unlikely that these desaturations were caused by the same physiological event, it is more probable that two independent desaturations occurred without any resaturation



(a) Graphical overview



(b) Overview of steps

Figure 6.1: Overview of the extraction of desaturations in the SpO_2 signal. (a) On top the MA filtered SpO_2 signal is plotted. In the middle, the derivative of this signal is shown, as well as its absolute value (solid line), on which peaks are detected (*). On the bottom, the start and end points of the desaturations and resaturations corresponding to these peaks are detected and linked together (o). As such desaturation events are extracted, containing a desaturation (D), a possible stable period (S), and a resaturation (R).

in between [95]. Therefore, these desaturations were seen as two separate events. Upward segments were linked in the same way to create complete resaturations.

An example of an interrupted desaturation with a plateau in between can be seen in Figure 6.1a, around the 60 s time point, 2 separate desaturation events were thus detected. The first event does not have a resaturation, whereas the desaturation part of the second event starts below the baseline SpO_2 value. The ability to detect nearby desaturation events with incomplete desaturation or resaturation, is the main advantage of using the derivative of the SpO_2 signal instead of just taking the extrema in the signal. This is very important, since these desaturation events could be detected as coming from two separate apneic events.

Finally, the start and end points of these desaturations and resaturations were extracted, using the extrema in the SpO_2 signal. Desaturation events were

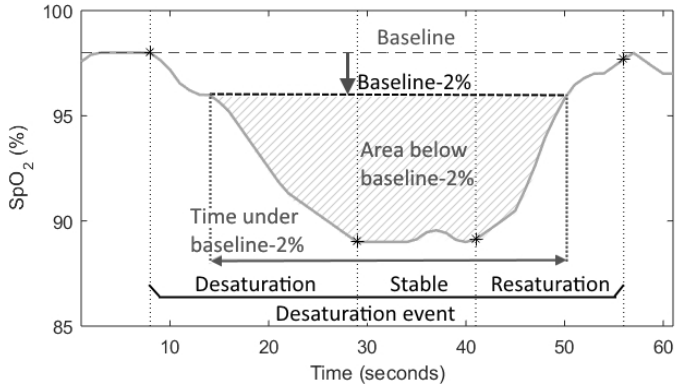


Figure 6.2: Extraction of desaturation severity features. The computation of area and time below 2 % under baseline are shown. Start and end points of the desaturation and resaturation are depicted with *.

retained if they had a drop in SpO_2 of at least 1 % and a total length, from the start of the desaturation to the end of the resaturation, shorter than 120 seconds, since this is typically the maximal duration of apneic events and their corresponding desaturations [72, 76].

Afterwards, a running SpO_2 baseline was computed as the 95th percentile value over the previous one minute recording that did not contain any detected artifacts or desaturations. A maximum baseline alteration of 1 % of oxygen saturation per second was allowed. When the computed baseline update had a change larger than 1 % with respect to the previous baseline value, the baseline was updated with only 1 % in the direction of the computed baseline update value.

6.2.3 Class label computation for the desaturation events

Sleep apnea scoring of PSG recordings is performed based on the respiratory signals; the start annotation of events represents the beginning of the flow limitation and the end annotation is placed when normal breathing is restored [17]. The start of the SpO_2 desaturation, however, has a delay of about 20-40 seconds with respect to the start of the apnea [95]. For each annotated apneic event, therefore, a desaturation event was sought within a window of 60 seconds after the start of the annotation. The apnea was linked to the closest desaturation in this window, which was not yet linked to a previous apnea. Sometimes apneas could not be linked to a desaturation, because there was no

drop in SpO_2. In this case, these apneas could not be detected by this algorithm. All remaining unlinked desaturations, were labeled as non-apneic for training and testing, these might for example be due to short breathing disturbances which do not meet the AASM 2012 criteria to be scored as apneic [17].

As such, all desaturation events from the SHHS and UZ Leuven data sets were labeled as being linked to an apnea or not. The Physionet data set, however, only has 1-minute annotations. Therefore, desaturation events could not be labeled individually, and thus this data set could only be used for testing with a 1-minute based performance evaluation.

6.2.4 Feature extraction

A total of 143 features were extracted from each detected desaturation event, and an overview of these features is given in Table 6.1. The features can be divided in simple time domain, desaturation severity, statistical and quasi-periodicity features.

Table 6.1: Overview of the extracted SpO_2 features.

#	Extracted features	Source
16	Amplitude (differences), lengths and slopes	
1	Area below baseline × total length	[76]
18	Area and length below baseline, baseline - 2/3/4 %, start SpO_2 level	[138]
50	Min, max, mean, median and variance of SpO_2 and its 1st, 2nd and 3rd order derivative	[72]
16	Deviation from min and max for median and mean of SpO_2 and its 1st, 2nd and 3rd order derivative	[72]
40	PRSA features, upward and downward slopes with fragment lengths of 10 and 20 s	[10]
2	1st peak of the autocorrelation and its relative amplitude	
4	Age, BMI, smoking status, gender	

Simple time-domain

A first group of parameters are simple time-domain features; amplitudes, amplitude differences, lengths and slopes were extracted from the desaturation and resaturation parts.

Desaturation severity

Next, desaturation severity parameters were computed, such as the obstruction severity, defined by Kulkas et al. as the area below baseline multiplied by the total length of the desaturation event [76]. In addition, the area and time below baseline or 2, 3, or 4 % under baseline during the event as introduced by Watanabe et al. and explained in Figure 6.2 were extracted [138]. This area and time were also computed with respect to the starting SpO_2 level of the desaturation event.

Statistical

The third group of features are the minimum, maximum, mean, median and variance of the SpO_2 signal and its first, second and third order derivative during the complete desaturation event, the desaturation part and the resaturation part. Additionally, the deviation from the means and medians by the minima and maxima were computed for the complete events, as proposed by Koley et al. [72].

Quasi-periodicity

The last group introduces novel features which capture the quasi-periodicities in the SpO_2 signal in a five-minute segment around the start of the desaturation event. As discussed in Chapter 3, apneic events rarely occur alone, there are often periods where apneic events follow in close succession as shown in Figure 3.3. As such, the desaturations occur one after the other, leading to periodicity in the SpO_2 signal. This thesis proposes to capture this periodicity using two methods: Phase Rectified Signal Averaging [10] and the autocorrelation function (ACF) of the SpO_2 signal.

The PRSA method searches for anchor points, which represent all time points where the signal goes downward (or upward) in the five-minute segments. Fragments of duration d are extracted around each anchor point, all these fragments are aligned and averaged. As such, an average curve of length d is

computed that represents the average desaturation (or resaturation) of the five-minute segment. From this curve, the capacity [68], the amplitude difference, overall slope, slope before and after the anchor point can be extracted as features. For the fragment duration d, 10 and 20 seconds were considered. Another way to detect the periodicity in the signal is by means of the autocorrelation function. The relative amplitude and position of the first peak in the ACF on the five-minute segments were extracted. The position of this peak represents the delay in between successive apneas, the relative amplitude gives a notion if a periodicity is indeed present.

From all 143 features, the logarithmic transformation was computed, which will also be included in the feature selection step. In addition, patient characteristics as age, BMI, smoking status and gender were considered. In total, 294 features were thus used as input for the feature selection described in the next section, from which 286 features were extracted from each SpO_2 signal and 8 from patient characteristics.

6.2.5 Feature selection and classification

In order to create a good classifier, the most discriminative feature set to separate desaturation events caused by an apneic event from those that are not, should be selected. An overview of the methodology used for the feature selection is given in Figure 6.3, and the different steps will be explained below. Both the feature selection and classifier training were based solely on the training set SHHS1 train, whereas the other data sets were used as independent test sets.

The first feature selection step was to delete highly correlated features to avoid multicollinearity in the classification system. Feature pairs with a correlation larger than 0.75 were investigated, in decreasing order of correlation. The feature with the lowest F-score was deleted, the other one was retained. In order to avoid overfitting of the feature selection to one patient, this F-score was computed using LOPO cross validation, and the overall F-score was then taken as the median of the LOPO F-scores. This resulted in a total of 68 retained features, with a correlation smaller than 0.75. The maximal correlation value of 0.75 was obtained as a trade-off between the number of retained features and the computational cost.

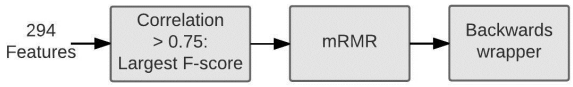


Figure 6.3: Overview of the different steps in the feature selection methodology.

The next step was to apply the mRMR algorithm to select the optimal feature set of 10 features. This number of 10 features was a trade-off between having all of the most informative features and reducing the computational cost of the backwards wrapper in the next stage, which was needed to remove any redundant features that might be left in the mRMR feature subset. This backwards wrapper was applied using RBF LS-SVM, using the `LS-SVMlab1.8` Toolbox [123] and 10-fold CV for parameter tuning. Since the subjects had on average around 400 desaturations per night, and there were 500 patients in our training set, the total training set consisted of around 200,000 data points. This was too large to be used for LS-SVM classification, as the kernel matrix of n x n should be computed with n the number of training data points. Therefore, a fixed-size LS-SVM classifier was trained on the 2000 most representative training data points, selected using the k-medoids algorithm on the features and class labels [134].

Using the selected feature set, classifiers were trained to decide if a desaturation was caused by an apneic event or not. The predictive value on the test sets was used to compare the different classifiers. Fixed-size LS-SVM classifiers using linear and RBF kernels, were compared against kNN [31], LDA [51] and RF [21]. The kNN classification was implemented using the Matlab function `fitcknn` and the optimal k was obtained with 10-fold CV. LDA was implemented using the `fitcdiscr` function, whereas the random forest classifier was implemented using `TreeBagger` with 100 bagged trees. All classifiers, except for the LS-SVMs, were computed both on the full training set and the reduced set of 2000 desaturations, in order to compare the obtained results with the LS-SVM classifiers.

After training, these classifiers were tested on the SHHS1 test, SHHS2, Physionet and UZ Leuven matched data sets. The performance was computed by comparing the class labels of the desaturation events with the predicted labels from the classifiers. For the Physionet data set, however, performance was computed by comparing 1-minute classification labels with the 1-minute annotations, as no event annotations were available. A 1-minute segment was classified apneic if a desaturation classified as caused by an apneic event started within the start of the 1-minute and 15 seconds after the end of the 1-minute interval.

6.2.6 Statistical analysis of classification performance

Classification performance was analyzed using the accuracy, sensitivity, specificity, positive predictive value, AUC and Cohen's Kappa value. The

performance parameters were computed for the point on the ROC curve that maximized the product of Se and Sp.

6.2.7 AHI estimation

Based on the number of desaturations classified as caused by an apneic event, an estimation could be made of the AHI for each patient. This was done by dividing the number of apneic classified events by the total recording time in hours. The annotated AHI, however, uses sleeping time instead of total recording time. This means that our computed AHI will be an underestimation of the annotated AHI. In addition, not all apneic events will result in a desaturation, and thus not all apneic events could be detected. In order to counteract these limitations, a robust linear regression was performed [61], using the `robustfit` Matlab function, between the annotated and computed AHI of the training set. This regression was then also applied to inflate the computed AHI of the test sets. Based on these regressed AHIs, classification of patients in OSA severity groups could be performed.

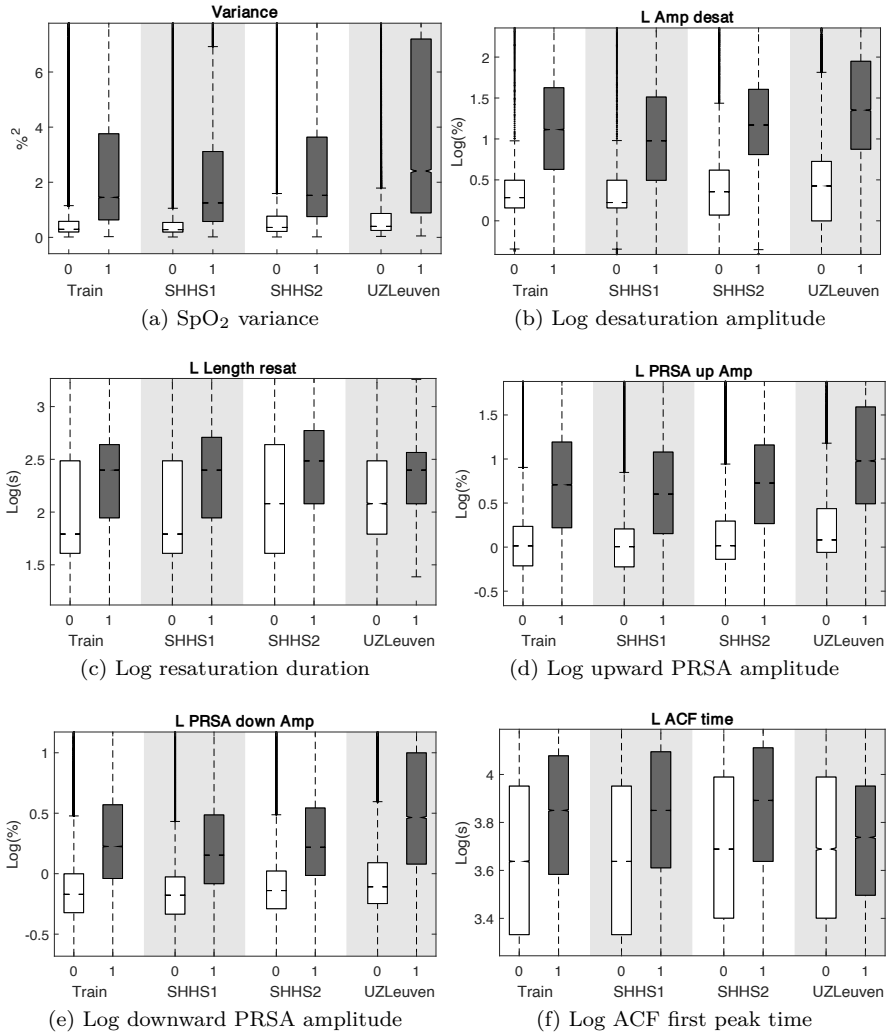
6.3 Results

The results of the apneic event detection algorithm on the training and test sets are discussed in this section. The selected features will be discussed and the performance of the different classifiers will be compared. The estimate of the AHI values based on the best performing classifier will be presented, as well as the OSA severity group classification.

6.3.1 Feature selection

The previously discussed feature selection techniques, applied on the training set SHHS1 train, resulted in the selection of six features. Three time domain features were selected: the amplitude of the desaturation, the variance of the SpO_2 signal and the duration of the resaturation. The three other features are novel quasi-periodicity features: the ACF time and the amplitude difference for upward and downward PRSA of 20 second fragments. All features, except for the SpO_2 variance, were selected in their logarithmically transformed form.

Boxplots of these features can be found in Figure 6.4 for desaturations linked to an apneic event (1) or not (0), and for each data set separately. A statistically significant difference was found between the apneic and non-apneic desaturation



(a) SpO$_2$ variance

(b) Log desaturation amplitude

(c) Log resaturation duration

(d) Log upward PRSA amplitude

(e) Log downward PRSA amplitude

(f) Log ACF first peak time

Figure 6.4: Selected features for desaturations linked to an apneic event (1) and not (0), for the SHHS and UZ Leuven matched data sets.

groups for all features and all data sets using the Wilcoxon rank sum test. Due to the large sample size, each group contains at least 25,000 samples, these differences might, however, represent a small shift in median values. When the boxplots of Figure 6.4 are studied, one can nevertheless see a clear difference in median values for all cases, except for the ACF time within the UZ Leuven

matched data set. Moreover, the inter quartile ranges have comparable values for the same group over the different data sets, and do not overlap between the apneic and non-apneic desaturations for four features (Figure 6.4a, 6.4b, 6.4d and 6.4e). No data is available for the Physionet data set, as no ground truth annotation of desaturations is available. The general trend is that desaturations produced by an apneic event have larger amplitude differences of desaturation and in PRSA curves, a longer resaturation time, more variance in the SpO_2 signal during the desaturation and a larger time until the first ACF peak.

6.3.2 Classification performance

Performance measures for classifying desaturation events caused by an apneic event or not are given in Table 6.2 for the different classifiers, averaged over all test data sets, using the number of subjects per test set as a weighting factor. The suffix *(2000)* represents the fact that these classifiers were trained only on the subsample of 2000 training desaturations and not on the complete training set. As can be seen from Table 6.2, the accuracy of the random forest classifier on the complete training set outperformed all the others. Its accuracy is at least 3 % higher. The AUC is similar for all data sets, except for Random Forest (2000) which had a lower AUC. As a result, the random forest classifier was used in the remainder of this work. When using the random forest classifier, a trade-off had to be made between the number of trees and the number of training samples used, due to computational limitations. Simulations have shown that it was better to use the full training set with a limited number of 100 trees.

Table 6.2: Overview of the averaged classification performance on the test data sets.

Classifier	Acc	Se	Sp	PPV	AUC	κ
Linear LS-SVM	79.0	75.8	79.9	54.6	85.3	49.1
RBF LS-SVM	78.7	76.6	79.3	54.1	85.4	48.9
k-NN	79.7	76.1	80.8	55.8	**86.0**	50.5
k-NN (2000)	79.5	76.1	80.4	55.4	85.7	50.0
LDA	78.7	**76.7**	79.3	54.1	85.3	48.9
LDA (2000)	79.3	75.2	80.5	55.1	85.3	49.5
Random Forest	**82.8**	64.3	**88.6**	**64.2**	85.4	**52.7**
Random Forest (2000)	77.8	74.4	78.7	52.7	83.7	46.5

Table 6.3: Overview of the averaged classification performance of the random forest classifier on each data set.

data set	Acc	Se	Sp	PPV	AUC	κ
SHHS1 train	82.8	64.6	89.3	68.4	85.7	54.9
SHHS1 test	83.3	60.8	90.2	65.4	85.0	52.2
SHHS2	82.0	70.7	85.7	61.1	86.1	53.5
UZ Leuven matched	78.1	75.5	80.8	79.5	85.4	56.3
Physionet	91.3	98.9	86.6	82.2	95.5	82.4

Table 6.3 shows separately the averaged performance parameters for each data set. Mean out-of-bag performance is given for the training set, which yields an accuracy of 82.8 %.

6.3.3 Estimation of the AHI

The AHI was computed as the number of desaturations caused by an apneic event divided by the total recording time. In Figure 6.5, on top, the computed AHI is plotted against the annotated AHI for each data set. As expected, the computed AHI underestimates the annotated AHI due to the fact that subjects are on average only sleeping 75.9 % of the recording and 11.5 % of the apneic events is on average not accompanied by a desaturation. The linear regression between the two AHIs was computed for each data set separately, resulting in good fits giving high R^2 coefficients of determination between 0.86 and 0.95. In order to ensure the independence of the test set, robust linear regression was computed on the training set and then applied to the test sets. This regression estimation of the AHI was used in the remainder of this study.

Figure 6.5 contains on the bottom, the Bland-Altman plots between the estimated and the annotated AHI for each of the data sets. These plots show an overestimation of the AHI for high AHIs, most clearly in the UZ Leuven matched and Physionet data sets. But to a lesser extent also in the SHHS test sets. The 95 % limits of agreement only slightly increase from the SHHS training set to the SHHS test sets, a limit of 10.9 is found for the training set and 14.8 and 14.7 respectively for the SHHS1 and SHHS2 test sets. The median absolute errors for the AHI estimate are for these three data sets, respectively, 1.19, 3.14 and 2.94. For the other two data sets, however, the limits of agreement double due to the overestimation of large AHIs.

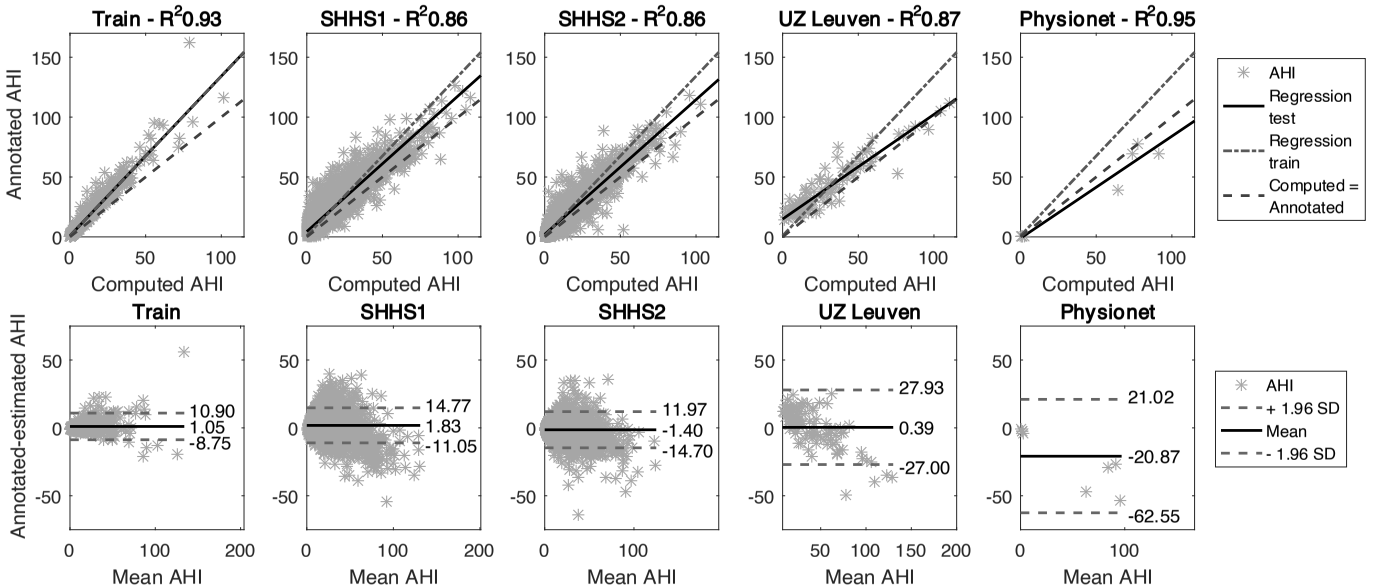


Figure 6.5: Top: scatter plot of the computed and annotated AHI for all data sets, R^2 values are given. Regressions based on the training and test data are also plotted. Bottom: Bland-Altman plot between the estimated and the annotated AHI.

6.3.4 OSA classification

Using the estimated AHI, the subjects were classified as having no (AHI<5), mild (5≤AHI<15), moderate (15≤AHI<30) or severe (AHI≥30) OSA. Figure 6.6 represents the confusion matrix of the estimated and true class labels for the SHHS data sets. Excellent classification performance was achieved on the training set, 83.2 % of the subjects was classified in the correct class, the other 16.8 % were classified in a neighboring class. Accuracies of 67.0 % and 71.9 % were achieved on, respectively, SHHS1 test and SHHS2, which is still quite good, but more subjects with mild and moderate OSA tended to get an underestimation of their OSA severity.

For OSA screening, an AHI threshold of 15 is often applied. To optimize this classification, the optimal point on the ROC curve was sought for classification of subjects with an AHI larger than 15 on the training set. A threshold for the estimated AHI of 14.52 was obtained. This is slightly lower than 15 and will counteract the underestimation of OSA severity in the test sets. Classification of OSA patients based on this threshold yielded an accuracy of 96 % on the training set and 87.5 % and 88.1 %, respectively, on the SHHS1 and SHHS2 test sets. OSA classification of the Physionet test set yielded 100 % accuracy, as there is a large gap between the AHIs of the two groups. OSA screening of the UZ Leuven matched data set is not useful as all subjects have an AHI larger than 15.

6.4 Discussion

6.4.1 Apnea event detection

This study proposed a method for apneic event detection based on desaturations in the SpO_2 signal which can be used to estimate the AHI and to screen for OSA. The oxygen desaturations were extracted by detecting upward and downward segments in the SpO_2 signal using its derivative. As such, all SpO_2 changes larger than 1 % should be detected, this was visually confirmed in a random subset of subjects. No annotations of desaturations were available. Hence, it was not possible to prove that all events of interest were detected for every subject. This could be seen as a limitation of the study.

From these desaturations, features were extracted and the most discriminating features were then selected. From the six selected features, three were novel quasi-periodicity features proving their usefulness for OSA detection. The feature selection included a backwards wrapper using the RBF LS-SVM classifier.

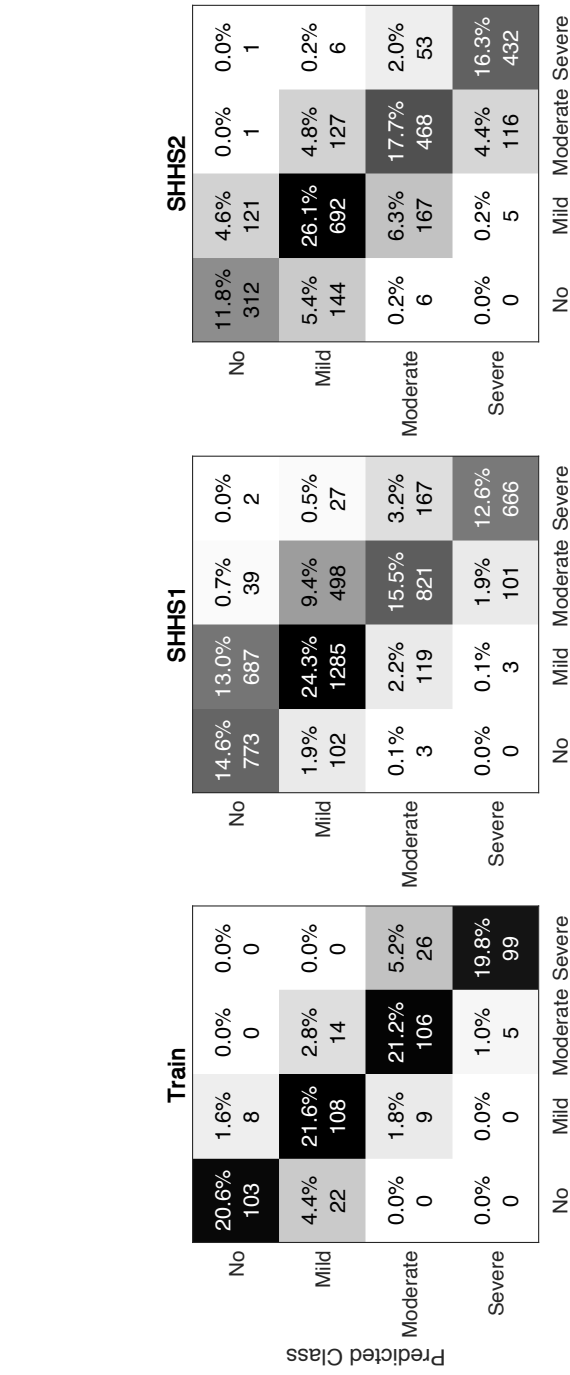


Figure 6.6: Confusion matrix for OSA severity classification based on the estimated AHI of the SHHS data sets.

This might have introduced a bias towards this classifier. However, only the final feature selection step included this classifier, which started from 10 features, selected independently from the classifier using the mRMR algorithm. Moreover, when comparing the test set performances of the different classifiers in Table 6.2, no bias towards this classifier was observed. Using only 2000 training samples the k-NN and LDA classifiers perform comparably. The random forest classifier, however, performs worse with 2000 training samples, but when the extra training samples were added this classifier outperformed the others.

The random forest classifier achieved an averaged accuracy of 82.8 %, which is lower than the performance achieved by Koley et al. [23], but that study was performed on other data sets and the method to annotate desaturations is not clearly given. Therefore, it is difficult to compare the results of both methods. The accuracy of 82.8 % is a good performance taking into account the fact that the algorithm was tested on a large population from different data sets. Moreover, the gold standard manual scoring of respiratory events suffers from intra and inter scorer variability, which has a negative impact on the degree of agreement [139].

When comparing the performances on the different data sets in Table 6.3, the Physionet data set stands out with the highest performance, namely an accuracy of 91.3 %. This can partly be explained by the 1-minute based evaluation. If a 1-minute segment contains 3 apneic events, and only 1 is detected by the algorithm, the segment will still be classified correctly. In contrast to the other data sets, where performance is evaluated for each desaturation separately and the previous scenario would only achieve a 33 % sensitivity.

Due to the 1-minute annotations, instead of individual apneic event annotations, this Physionet data might be seen as not perfectly fitting this study. The Apnea-ECG data set is, however, a publicly available and heavily studied data set, which makes it possible to compare our results with methods from the literature on the assessment of OSA severity. This comparison showed that our results were similar to those from the literature. An accuracy of 91.3 % was achieved in this study, whereas the SpO$_2$-based method proposed by Burgos et al. [23] yielded a performance of 93.0 % on the same data set. Despite the fact that the classifier in this study was not specifically trained on the Physionet data set or on 1-minute segment classification as is the case in the study of Burgos et al. [23]. SpO$_2$-based algorithms seem to outperform ECG-based algorithms on this data set, which achieve accuracies between 85 % and 90 % on the full Apnea-ECG database [33, 135].

The SHHS test sets have good performances, comparable to the one achieved in the training set. An accuracy of 83.3 % and 82.0 % is achieved for those data sets, however, the sensitivity is rather low. Figure 6.7 shows that this

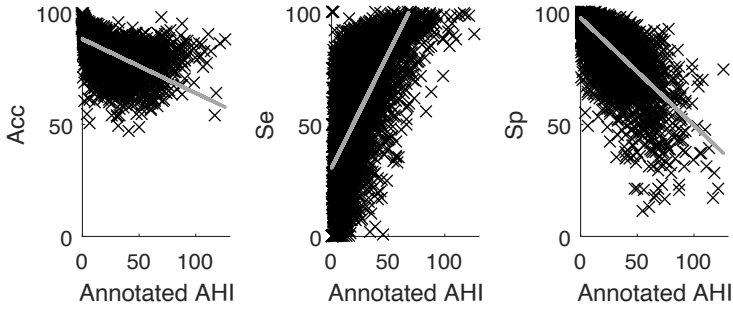


Figure 6.7: Accuracy, Sensitivity and Specificity plotted for each patient of the test sets in function of the patient's AHI with the linear regression line plotted on top. Accuracy and Specificity decrease with increasing AHI, whereas Sensitivity increases.

low sensitivity is mainly due to subjects with a low AHI. The Acc, Se and Sp are plotted for each test patient in function of the AHI. A decreasing trend can be found for the accuracy and the specificity if the AHI increases, whereas the sensitivity increases. In the SHHS 1 and 2 test data sets, only 17 % of the patients have severe OSA (AHI$\geq$30), most patients thus have a lower sensitivity and a higher specificity.

Moreover, when looking at the number of desaturations per hour of recording and the percentage of these desaturations that are labeled apneic, which are given in Table 6.4, it can be seen that subjects with larger AHI have more desaturations per hour and a larger percentage of these desaturations are apneic. To avoid a bias in the classifier, it thus might be a good idea to resample the training set to obtain the same number of apneic and non-apneic desaturations per subject. This was not done and is a limitation of the study.

Another important point is the fact that approximately 73 % of the events in the SHHS test sets are hypopneas, which are harder to detect than apneas. If only the apneas would be considered, the sensitivities would increase to, respectively, 77.8 % and 83.9 %. The remaining difference in performance between the SHHS1 and SHHS2 test sets might be caused by the age difference and the further progression of the OSA in between the two polysomnographies.

The accuracy of the UZ Leuven matched data set is 78.1 % which is slightly worse than the SHHS data sets. This is probably due to the fact that SHHS scoring of respiratory events is slightly different from the AASM 2012 [17] rules used in the UZ Leuven data sets. In the SHHS scoring, apneas are already scored when a reduction of respiration of 75 % occurs instead of the standard 90 % nowadays. The study population in the UZ Leuven data set also differs. Whereas the

Table 6.4: Number of detected desaturations per hour of recording and the percentage of these desaturations that is labeled apneic for each data set and OSA severity category (C=control, Mi=mild, Mo=moderate, S=Severe OSA).

	Desaturations/hour				% apneic			
data set	C	Mi	Mo	S	C	Mi	Mo	S
SHHS1 train	35	45	56	71	6	15	27	45
SHHS1 test	35	46	57	72	6	15	26	44
SHHS2	35	46	54	70	6	15	27	43
UZ Leuven matched	-	-	50	74	-	-	34	57
Physionet	42.1	-	-	88	-	-	-	-

other studies are investigating the general population, the UZ Leuven recordings are conducted on a patient population with identified OSA symptoms. This difference probably also explains the higher percentage of apneic desaturations for UZ Leuven subjects in Table 6.4. Moreover, all subjects had an AHI larger than 15, so moderate to severe OSA. When taking into account the tendency seen in Figure 6.7, one can conclude that in these categories the accuracy and specificity will be lower but a higher sensitivity is seen. This data set also contains a high number of hypopneas (66 %), hence, if only the apneas are taken into account, the sensitivity for the UZ Leuven data set raises to 92.6 %.

These results are obtained without using any form of sleep staging, some of the detected desaturations, however, occur while the patient is awake. When desaturations during awake are deleted from the analysis, the performance only slightly increases, by 0.5 % on average. As the added value of adding sleep staging is minimal, this was not further investigated.

6.4.2 AHI computation and OSA screening

From the detected apneic events, the AHI was computed using robust regression on the training set. As can be seen from Figure 6.5, this regression on the training set has for all test sets, a steeper slope than the regression on the test set. For the SHHS data sets the deviation between the two regressions is limited, but more prone for higher AHIs. This could be due to the under representation of severe OSA patients in the test sets. For the UZ Leuven data set, the number of apneic events not associated with a desaturation is much lower than the average, only 5.7 %. This causes the regression on this data set to be much closer to the identity line. Therefore, the inflation of the AHI due

to the training regression caused a large overestimation of the AHI, mostly in severe OSA patients. This difference in the number of apneic events associated with desaturations might be due to the fact that this data set was recorded on a patient population with OSA related complaints. It might be possible that these subjects are more prone to more severe apneic events associated with desaturations than subjects with a high AHI without any symptoms. Similar results are seen on the Physionet data set, although no information is available on the individual apneic events, only 5.3 % of the apneic minutes do not contain any desaturation.

This AHI regression difference between the separate data sets indicates that it might be useful to create different AHI regression estimates dependent on the data set at hand. This could for example be done by retraining the regression based on a subset of the population at hand. Or in case of a clinical trial, where no ground truth is available, the regression can be based on a similar data set. For example, if the focus of the clinical trial is on a population with OSA symptoms, AHI regression based on the clinical UZ Leuven data set should be preferred compared to the SHHS, which is a general population data set. Another option is to try to identify patient markers that could be responsible for altering this regression, which should be further investigated.

From the computed AHI, OSA severity classification of subjects was conducted, which achieved good results on the SHHS data sets. The results of the OSA subject classification with an AHI threshold of 15 on the SHHS2 data set are compared with methods from the literature [28, 113, 116, 137], as computed by Rolón et al. on a subset of 995 subjects of the SHHS2 data set in Table 6.5. Additionally, the method by Morales et al. [93] was applied on the SHHS2 data set and the results are included in the table as well. As observed, our methodology outperforms all other methods on this data set. The accuracy

Table 6.5: Comparison of OSA patient classification performance for AHI >15 on SHHS2.

Method	Acc	Se	Sp	AUC
This study	**88.08**	**90.12**	86.39	**0.953**
Morales et al. [93]	85.28	84.75	85.81	0.936
Rolón et al. [113]	85.78	85.65	85.92	0.937
Schlotthauer et al. [116]	85.02	84.11	85.94	0.922
Vázquez et al. [137]	84.17	80.84	**87.50**	0.909
Chiner et al. [28]	77.15	78.12	76.17	0.795

improves by more than 2 % to 88.08 % and the AUC also improves to 0.953. These results are even more remarkable since no data of the SHHS2 data set was used for training of this classifier. Only 500 subjects of the similarly recorded SHHS1 data set were used for training. But, 223 SHHS1 recordings of SHHS2 subjects were included in the training set. This means that 8.4 % of the subjects in this data set are not independent of the training set. The other 91.6 % of the subjects are, however, independent of the training set. It was investigated if the inclusion of these previous recordings of the same subject improved the results. This, however, did not show any positive bias towards subjects included in the training set, these subjects only had a severity classification accuracy of 85.2 % in SHHS2, with a Se, Sp and AUC of respectively 84.7 %, 85.6 % and 95.0 %.

Since most of the used features for classification of desaturations are linked with desaturation severity, the added value of this method compared to the conventional ODI 3 score was investigated. The ODI 3 parameter represents the number of oxygen desaturations of at least 3 % with respect to baseline per hour of sleep. Unfortunately this ODI 3 parameter is not available for the used data sets. As an alternative, however, the desaturation classification was repeated using only the percentage of desaturation with respect to baseline as a feature. A cut-off value of 3 % was used to mimic the ODI 3 parameter. This resulted in an averaged classification accuracy of 79.5 % on the test sets. Our method has a 3.3 % increase in performance compared to this conventional parameter. Moreover, the ODI 3 method has a very low averaged sensitivity of 41.6 %. Our method has an increase in sensitivity of more than 20 %, thus extra apneic events are detected, without compromising the specificity of the classifier. This also results in a lower OSA severity classification. This is to say, when only the desaturation from baseline is considered, accuracies of 83.5 and 81.9 are obtained, respectively on the SHHS1 and SHHS2 test sets. The computations were repeated for cut-off values of 2 and 4 %, but the results did not improve. Moreover, the use of the conventional ODI parameters has as drawback that no standard definitions of desaturation and baseline are available.

The results of the proposed processing methods show that OSA screening based on SpO_2 signals can achieve high performances. Moreover, the methods are fully automated, can handle noisy signal segments, do not need any sleep scoring and use only simple features extracted from the desaturations, therefore, they could easily be implemented in a wearable device for screening. Since the screening is, however, only based on the SpO_2 signal, sleep staging and detection of apneic events without desaturations will be lost compared to the full PSG.

The presented results could still be improved if extra features are considered, for example non-linear or frequency features [5, 93]. This, however, probably will lead to an increase in computational cost, which is a drawback for wearable systems. Moreover, the considered data segments are too short for

the computation of nonlinear parameters. A multimodal apneic event detection algorithm could also be created by for example adding information from the PPG signal, which is already available from the pulse oximeter. This will be investigated in Chapter 9. Further investigation into the proposed method is needed to see if a differentiation can be made between the different types of apneic events. Features should be sought which can differentiate central from obstructive events and hypopneas from apneas, as this gives clinicians crucial information on how to treat the OSA.

The next section will give an overview of the additional experiments that were performed after the publication of this study. These include the application of the RF-based feature selector, the retraining of the algorithm based on clinical data from the UZ Leuven data sets, and the use of the mICS model for OSA severity classification. All these experiments were performed on the UZ Leuven OSA+ data set instead of the UZ Leuven matched data set, as this set contains recordings with a full range of AHI values.

6.5 Additional experiments

6.5.1 Application of RF-based feature selector

The aim of this experiment was to improve the selection of the optimal feature set, used to predict whether an oxygen desaturation was linked to an apneic event or not, by implementing the proposed RF-based feature selection algorithm. More information on this novel feature selection algorithm can be found in Section 5.3.

The RF-based feature selection algorithm was applied to the feature set of 286 features extracted from oxygen desaturation of the SHHS train data set. The RF model trained on the obtained feature subset was then applied to the UZ Leuven OSA+ data set. The same classification and AHI estimation flow was followed as explained in Section 6.2.

Results

The RF-based feature selector chose the optimal number of features based on the kappa correlation performance parameter, which is shown in Figure 6.8. The optimal feature set was obtained when 283 features were removed, or in other words, when 3 features were retained: the Log amplitude of desaturation, the

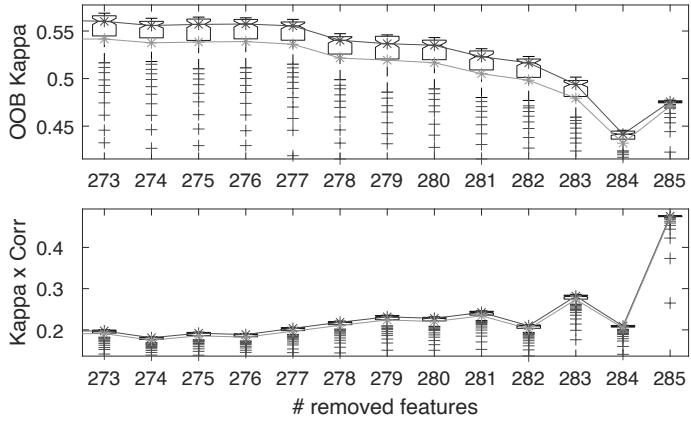


Figure 6.8: Overview of the classifier performance values, according to the number of features that were removed. On top, the OOB Cohen's kappa values per tree are plotted. On the bottom these values were multiplied with one minus the mean correlation coefficient in the feature set.

Table 6.6: Overview of the desaturation classification performance when using the pruned feature set obtained using the RF-based feature selection method.

	Data set	Acc	Se	Sp	AUC	κ
RF-Based feature selection 3 features	SHHS1 train	80.8	63.8	86.9	83.2	50.6
	SHHS1 test	81.6	60.0	88.2	82.2	48.4
	SHHS2	80.6	68.2	84.5	83.7	49.9
	UZ Leuven	76.6	68.9	80.9	81.6	49.6
Original model 6 features	UZ Leuven	79.6	71.4	84.2	84.9	55.7

Log downward PRSA amplitude and the variance of the second order derivative of the SpO_2 signal during desaturation.

The performance results of the desaturation classification for the different data sets can be found in Table 6.6. The table also includes the results obtained with the original feature set including 6 features for the UZ Leuven OSA+ data set. For the SHHS data sets, these results can be found in Table 6.3.

The results for the OSA screening according to AHI cut-off points of 5, 15 or 30 apneic events per hour of sleep are given in Table 6.7. The results of this

Table 6.7: Overview of the apnea screening performance, when taking an AHI cut-off of 5, 15 or 30 for the different data sets, using the original and RF-based feature set.

	Original model 6 features				RF-based feature selection 3 features			
	SHHS 1 train	SHHS 1 test	SHHS 2	UZ Leuven	SHHS 1 train	SHHS 1 test	SHHS 2	UZ Leuven
Acc5	96.2	83.0	89.7	81.2	89.2	84.4	89.8	81.2
Se5	97.1	81.3	93.1	79.8	90.1	84.2	93.4	81.4
Sp5	93.6	91.3	73.6	92.0	86.4	85.2	72.7	80.0
AUC5	99.1	93.1	94.2	93.1	93.9	92.4	93.8	90.3
Acc15	96.0	87.5	88.1	85.3	91.0	87.4	84.7	82.6
Se15	95.6	77.4	90.1	79.5	90.4	85.8	92.9	80.3
Sp15	96.4	95.4	86.4	93.4	91.6	88.7	77.8	85.7
AUC15	99.2	95.0	95.3	93.6	96.7	94.5	94.4	91.7
Acc30	93.0	93.1	90.3	91.7	91.8	93.4	91.0	89.0
Se30	83.2	84.6	92.5	89.2	80.8	85.9	90.5	86.5
Sp30	96.3	94.7	89.9	93.1	95.5	94.9	91.2	90.3
AUC30	97.1	96.9	97.0	97.9	96.4	96.8	96.6	96.6

experiment are again compared with those from the original 6-feature model.

Discussion

The proposed feature selection algorithm resulted in the selection of 3 features compared to the 6 features obtained before. Therefore, the number of features was reduced to half, whereas, this only resulted in a minor decrease in classification performance, see Table 6.6. Moreover, the clinical relevance of this drop in performance can be assessed by looking at the OSA screening parameters in Table 6.7, only slight changes in AUC were detected, while now only half of the features need to be computed. This greatly increases the interpretability of the classifier, the 3 features can now be visualized in a 3D

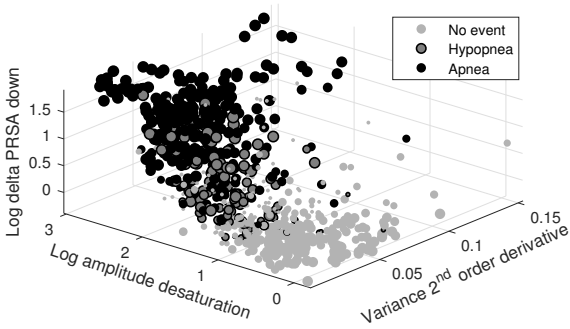


Figure 6.9: 3D plot of the selected oximetry features for one patient.

plot, as is shown in Figure 6.9 for one patient. This can help to define why certain desaturations or patients could not be classified correctly.

The RF feature selector uses the Cohen's kappa value as performance measure, which is very useful in unbalanced multiclass classification problems. The stability of the Cohen's kappa value over data sets with different characteristics can be noticed in Table 6.6. While the accuracy varies between the SHHS and UZ Leuven data sets, the kappa values remain stable. Moreover, the patient independent validation by selecting the OOB samples per patient is important in many clinical problems. This adaptation resulted in a more realistic estimation of the validation performance of the training set. This can be observed in the OSA screening results in Table 6.7, the training AUCs were around 99%, whereas they are now more in line with the test performances.

The algorithm was tested on the UZ Leuven OSA+ data set, which contains patients with various AHI levels. This enables the evaluation of the OSA screening performance on this clinical data set containing patients coming to the sleep laboratory for diagnosis of sleep apnea. The results in Table 6.7 show that the screening performance is quite stable between the general population train and test sets of the SHHS, and the clinical UZ Leuven data set. For an AHI cut-off value of 15 an AUC of 91.7-94.5 % can be achieved with an accuracy of 82.6-87.4 %, when shifting the cut-off to 30, an AUC of 96.6 % is obtained with an accuracy of about 91 %.

Preliminary conclusion

When applying the RF-based feature selection for OSA classification, an optimal feature set of 3 parameters was selected out of 286. This was half of the 6 features

that were obtained before. This feature reduction resulted in an improved interpretability of our model, but also a slight decrease in performance, without affecting the clinical screening performance. These results were presented at the 2019 EMBC conference [40].

6.5.2 Retraining classifiers on UZ Leuven data sets

As mentioned before, the SHHS data set might not be ideal to train an OSA screening algorithm intended to be used on clinical data sets. Therefore, it was tested if the classification performance of this algorithm on a clinical data set could be improved by retraining the model on data of a clinical population. To this end, the UZ Leuven OSA+ data set was used. From this data set, 100 subjects were selected as training set, based on patient demographics. The other 118 will be used as independent test set.

Methods

In total, four new models were retrained. An overview of the training sets used to train the desaturation classification and AHI regression for these four models is given in Table 6.8. In the first model, *OSA+_full*, the whole algorithm was retrained only using the 100 training subjects of the UZ Leuven OSA+ data. The second model, *OSA+_AHI*, used the same training data, but only retrained the AHI regression and OSA severity classification. To decide if a desaturation is apneic or not, the classifier trained on the SHHS data was kept. The third and fourth model, *mixed_full* and *mixed_AHI*, performed the same classifier

Table 6.8: Overview of the training data used to retrain models on the UZ Leuven data sets

	Classification of desaturations			AHI regression		
Training set	500 SHHS	100 OSA+	500 SHHS + 100 OSA+	500 SHHS	100 OSA+	500 SHHS + 100 OSA+
Original	X			X		
OSA+_full		X			X	
OSA_AHI	X				X	
mixed_full			X			X
mixed_AHI	X					X

retraining, but took as new training set the combination of the 500 SHHS train subjects and the 100 UZ Leuven OSA+ training subjects. This analysis was performed on the original model including the full set of six features.

Classification of desaturations as apneic or not

In Table 6.9, the obtained desaturation classification performances on the UZ Leuven OSA+ test set are presented. The original model obtained a similar AUC and κ score as on the UZ Leuven matched data set in Table 6.3. The sensitivity dropped a bit, however, while the specificity increased. This can be explained by the trends seen in Figure 6.7, as the UZ Leuven OSA+ test set also contains subjects with low AHIs. When the classifier was retrained only on the UZ Leuven patients, no increase in AUC could be observed. Moreover, a clear drop in sensitivity occurred. But when a mixture of both data sets was taken as training set, an increase in classification performance was obtained on the independent test set. For the *_AHI* models, the performances were the same as the original model, as this part of the model was not changed in these retrained models.

AHI estimation

Next, the AHI regression was retrained, and subjects were classified as having an AHI larger than 15 or not. The results of this classification on the test set for the retrained models can be found in Table 6.10. The test set contained 73 subjects with an AHI larger than 15, and 45 with an AHI lower than 15. All models obtained similar AUCs of about 93.5 %, which is comparable to the results obtained on the SHHS 2 test set. When focusing on the accuracy, however, a clear difference between the models' performances can be seen, in which the retrained *mixed_full* model outperforms the other models, obtaining an accuracy, sensitivity and specificity of about 89 %. The other retrained models obtain poorer accuracies than the original model, mainly caused by

Table 6.9: Overview of the desaturation classification performance for the retrained model on UZ Leuven OSA+ test.

Model	Acc	Se	Sp	PPV	AUC	κ
Original	79.6	73.9	83.1	73.0	85.6	56.9
OSA+_full	**80.0**	67.7	**87.6**	**77.2**	85.7	56.6
mixed_full	79.9	**74.3**	83.4	73.4	**86.0**	**57.5**

Table 6.10: Overview of OSA patient classification performance of retrained models for AHI >15 on UZ Leuven OSA+.

Model	Acc	Se	Sp	AUC
Original	83.90	76.71	95.56	93.5
OSA+_AHI	81.36	69.86	**100.00**	93.5
mixed_AHI	81.36	72.60	95.56	93.5
OSA+_full	82.20	72.60	97.78	93.5
mixed_full	**88.98**	**89.04**	88.89	**93.6**

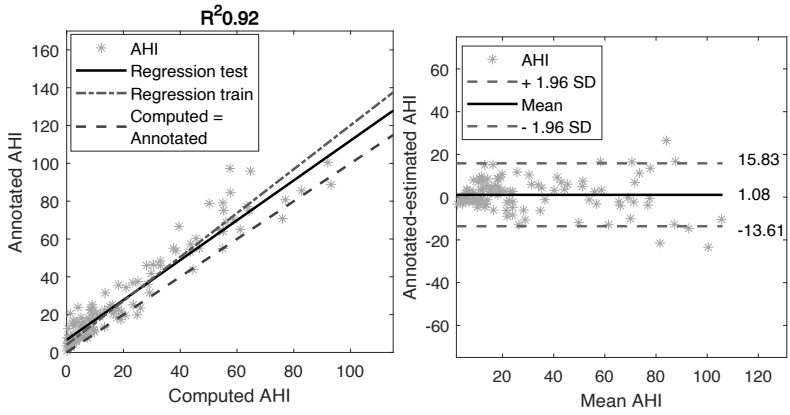
an imbalance between sensitivity and specificity. This imbalance is partially explained by the fact that the optimal classification threshold obtained based on the training data, is not ideal for the test set.

Final model

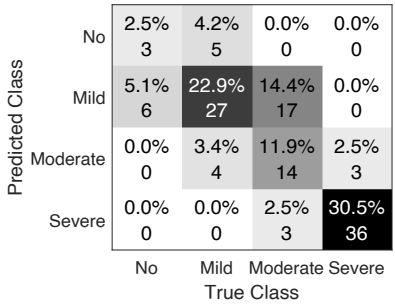
In Figure 6.10, more results are given for the AHI estimation and OSA severity classification using the *mixed_full* model on the UZ Leuven OSA+ test set. In the scatter plot of Figure 6.10a, the regression line of the training and test set align very well for AHIs lower than 80. Moreover, The Bland-Altman plot in Figure 6.10b shows similar limits of agreement as were seen for the SHHS test sets in Figure 6.5. When 4-class OSA severity classification was performed, 67.8 % of the subjects were classified in the correct class and misclassified subjects were predicted to belong to a neighboring class, as can be seen in Figure 6.10c. Most of the errors come from moderate OSA subjects that were classified as having mild OSA.

Preliminary conclusion

From these results, it can thus be concluded that retraining the model using a training set including clinical data can improve the performance on the clinical data sets, both for the desaturation classification and AHI regression. However, this was only the case when the clinical training set was added to the original training set of 500 subjects. This might be due to the fact that a training set of 100 subjects is not large enough for the model to obtain all necessary information. The SHHS training data might have been especially interesting for the model to learn from data recordings with very low AHIs, which are rare in the clinical data set.



(a) Computed vs annotated AHI (b) Bland-Altman plot



(c) Confusion matrix

Figure 6.10: Overview of the AHI estimation and OSA severity classification performance of the *mixed_full* retrained model on the UZ Leuven OSA+ test set.

6.5.3 OSA screening using a mICS model

In the previous experiment, it was shown that the AHI estimation on clinical subjects could be improved when the model was retrained on a training set including subjects with more similar characteristics. In this experiment, the effect of including patient demographics and clinically used SpO_2 parameters on the AHI estimation was investigated. In the study of Behar et al. [12], it was shown that including these parameters could increase the AHI estimation performance.

A multilevel ICS model was trained to predict the 4 OSA severity levels based on the number of desaturations classified as apneic per hour of recording, and

additional demographic and clinical parameters. This model will replace the robust linear regression step from the original algorithm explained in Section 6.2.

Methods

For this experiment, three training sets were considered, similar as in the previous experiment. The training set included either the 500 subjects of SHHS1 train, the 100 training subjects of UZ Leuven OSA+, or the combination of both. The obtained classifier was then tested on the remaining 118 subjects of UZ Leuven OSA+, SHHS1 test, and SHHS2.

The estimated AHI, computed by the original OSA screening algorithm as the number of oxygen desaturations classified as apneic per hour of recording, is considered as an input feature to the mICS model. Additionally, the age, BMI and gender were included for each subject as demographic parameters. Moreover, a parameter was included which tracked whether the measurement was recorded from a clinical population (UZ Leuven), or the general population (SHHS). Finally, also clinical SpO_2 scores were added, including the ODI 2, 3 and 4, the T90, the mean oxygenation level, the nadir SpO_2, and the 1, 5, and 10 % percentile values of the SpO_2 signal.

This feature set was taken as input to the multilevel ICS model which was trained to predict the 4 OSA severity categories: no, mild, moderate and severe OSA. The model was retrained with and without taking into account parameter interactions and with and without elastic net parameter pre-selection. Different values for the AUC cut-off hyperparameter were also tested.

Results

The returned optimal mICS model always only took into account the estimated AHI, even when all the discussed training parameters were adapted. Only, when an AUC cut-off value of 0.999 was taken, the model would include other inputs. But this model would then overfit to the training set and would not be useful in practice. When the estimated AHI was removed from the feature set, mICS models which only included the ODI3 were obtained. It is thus clear that the SpO_2-based alternatives for the AHI dominate the model, and that the patient demographic features do not provide added value to them.

The best classification was obtained when the mixed training set was used. This resulted in a model which classified subjects as having no apnea if they had an estimated AHI lower than 3, mild apnea when this value was between 3

and 8.75, moderate apnea with an estimated AHI between 8.75 and 19.54 and severe apnea with an estimated AHI above 19.54. These thresholds are below the clinical cut-off points, as not all apneic events are linked to a desaturation, and since the recording time is an over-estimation of the time asleep.

This model classified 66.7 % of the SHHS1 test set in the correct severity class, 60.1 % of the SHHS2 set and 65.5 % of the OSA+ test set. These accuracies are similar but slightly lower than what was obtained by the robust linear regression in the original algorithm.

Preliminary conclusion

The proposed demographic and clinical SpO_2 parameters did not show to have an added value to estimate the AHI severity category. As only one parameter was included in the final model, there is no advantage in using the mICS model compared to the robust linear regression for this use case. Applications in which more features contribute to the classification estimate are needed in order to prove the benefit of the mICS model. In Chapter 12 the mICS model was applied for cardiovascular phenotyping of OSA subjects, in which it was shown to be useful.

6.6 Conclusion

A methodology for the automatic detection of respiratory events using SpO_2 signals based on the detection and classification of desaturations caused by apneic events was developed. A random forest classifier based on six desaturation severity and SpO_2 periodicity features achieved the best performance for classifying desaturations. An averaged accuracy of 82.8 % was achieved over different independent test sets. Robust linear regression was used to estimate the AHI from the number of desaturations classified as apneic. It was shown that the optimal regression differed between the data sets, to estimate the AHI, the data set characteristics should be taken into account. The estimated AHI was used for OSA severity classification and OSA screening with an AHI threshold of 15. This screening achieved an accuracy of 88 % on the SHHS2 data set, outperforming all SpO_2-based methods from the literature tested on the SHHS2 data set.

Additional experiments have shown that the feature set used to classify desaturations could be further pruned from 6 to 3 features with minimal loss of performance. Moreover, the apnea detection and OSA screening performances on clinical data sets could be improved when recordings of clinical subjects were

added to the training set. Retraining the algorithm on this new training set resulted in an OSA screening performance of 89 % for the UZ Leuven OSA+ data set. Adding patient demographics and clinical SpO_2 scores to enhance the AHI estimate using mICS models did not improve the results.

These results show that this computationally cheap methodology can be very useful in an OSA home monitoring system based on pulse oximetry. The developed algorithm was included in a graphical user interface (GUI) implemented in Python, which is described in Appendix A.

Chapter 7

ECG-based sleep apnea detection

This chapter will present an algorithm for the automatic detection of apneic events from single-lead ECG signals. The algorithm builds further on the work presented by Varon et al. [135], but will introduce additional cardiorespiratory measures and an alternative EDR estimate. The algorithm was designed, in such way that it also can be applied on (capacitively-coupled) ECG measures combined with an unobtrusive respiratory measurement.

This work was performed in collaboration with Imec, and the methods presented in this chapter will be published as a part of: ***Deviaene, M.****, Castro, I. D., Borzée, P., Patel, A., Torfs, T., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Capacitively-coupled ECG and respiration for the detection of sleep apnea, submitted to Physiological Measurement.*

In the next sections, the used data sets will be discussed, followed with the data preprocessing including heart rate and EDR extraction. Subsequently, the feature extraction, selection and classification will be explained. In Section 7.3 the results will be presented, which will be further discussed in Section 7.4.

7.1 Materials

For this study, single-lead ECG signals extracted from PSGs of 3 UZ Leuven data sets were used. The OSA+ data set was used to train the classifier, and the Imec and Varon data set were used as test sets. More information on these data sets, and an overview of the patient demographics of each data set can be found in Chapter 2.

The ECG signals of the OSA+ and Imec data set were sampled at 500 Hz, whereas the Varon data set was sampled at 200 Hz. All signals were upsampled to a uniform sampling frequency of 512 Hz. This sampling frequency was chosen to match the sampling frequency of the ccECG, which will be used in Chapter 10.

7.2 Methods

Before analyzing the signals, the data was preprocessed in order to remove noise and artefacts. Then, the EDR and RR-interval time series were extracted. From these signals, an apnea detection classifier was developed, features were computed of which the most relevant ones were selected and fed to a classifier. This model was then tested on the Varon and Imec data sets.

7.2.1 Data preprocessing

An overview of the preprocessing steps needed to prepare the ECG signals for feature extraction is given in Figure 7.1. These steps will be discussed in more detail below. After signal prefiltering, the signals were split into 1-minute windows and the further analysis was performed per window.

ECG prefiltering

The 50 Hz power line interference was removed from the ECG signals using a notch filter. Next, a band pass Butterworth filter was applied using cut-off frequencies of 0.67 and 40 Hz [25, 26], and the mean was removed from the signal. The used low-pass cut-off frequency of 40 Hz is lower than what is recommended by the task force of heart rate variability [82]. However, this algorithm was developed such that it also could be applied to unobtrusive ECG measurements. When this upper limit is increased for unobtrusive ECG measurement systems,

a significant increase in noise levels is observed. Therefore, the IEC 60601-2-47 standard for ambulatory ECG systems was followed to define the filtering band.

Check for signal inversion

A check for signal inversion needed to be implemented since for the PSG signals the placement of the ECG electrodes was sometimes adapted. When the absolute value of the 1st percentile value of the ECG signal amplitude was larger than the 99th percentile value, lead inversion was assumed and the signal was inverted.

Artefact detection

For each 1-minute window, a signal quality score was computed. This was done based on the autocorrelation function of the ECG [90]. Segments with a quality score lower than 90 were discarded.

RR-interval extraction

The R-peaks were extracted from the ECG segments using the algorithm discussed in [135], which makes use of the upper and lower envelopes to flatten the ECG signal. Based on the location of the detected R-peaks, the RR-interval time series, or tachogram was derived.

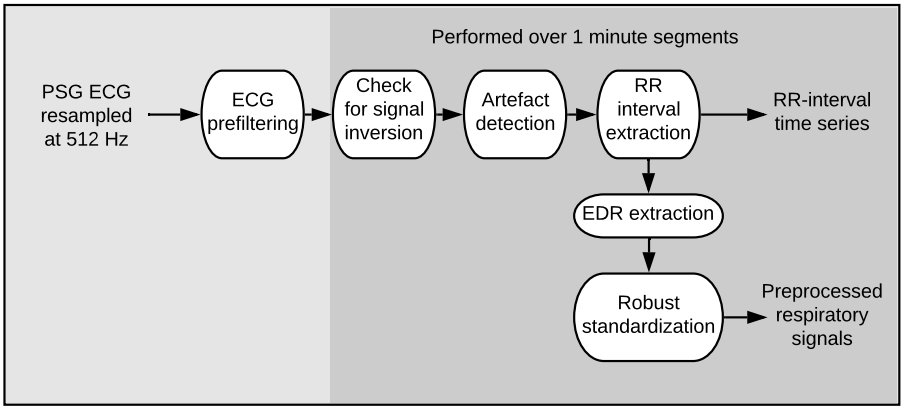


Figure 7.1: Overview of the applied ECG signal preprocessing methods.

EDR extraction

From each 1-minute window, where at least 30 R-peaks were detected, the EDR was computed. Two EDR algorithms were compared: the R-wave amplitude (Ramp) and the QRS slope range (SR) [73]. These were extracted as described in [132].

Robust standardization of the EDR signals

In order to ensure that in the future the EDR features could easily be replaced by features extracted from unobtrusive respiratory measurements, a robust standardization was applied to the EDRs. The data was rescaled such that the 10th and the 90th percentile values of the entire signal corresponded to values of, respectively, -1 and 1.

Conversion to 1-minute labels

As a last preprocessing step, the continuous apnea annotations of the PSG were converted to 1-minute labels. Whenever a 1-minute segment overlapped with an annotated event (including both apneas and hypopneas), the segment was labeled as apneic. These 1-minute labels were used as reference to train and test the classifier.

An overview of the available number of clean hours of data can be found in Table 7.1.

Table 7.1: Overview of the number of hours of clean ECG data.

	Hours of clean data	Total hours of recording	
Data set	ECG	Normal	Apneic
OSA+	1547	1381	609
Imec	193	172	83
Varon	82	64	35

7.2.2 Training apnea detection classifier

The training workflow to develop the apnea detection classifier is depicted in Figure 7.2 (top). Afterwards, the classifier was tested on the Imec and Varon data set for both EDR methods.

Additional experiments were conducted in order to explain the obtained results (see Figure 7.2). The different steps in the development of the classifier are explained below.

Feature extraction

ECG features: A total of 27 features were extracted from the ECG segments. The mean, standard deviation (std) and kurtosis of the RR-intervals were extracted, as well as the SDSD and the RMSSD, the first five serial correlation coefficients of the RR-intervals and the Allan factor (on time scales of 5, 10, 15, 20, 25 and 30 s) [33]. Additionally, the $\text{QRS}_{\text{morph}}$ feature quantifies changes in the QRS morphology. This feature was computed by applying principal component analysis on a matrix composed of the aligned QRS complexes. $\text{QRS}_{\text{morph}}$ was taken as the percentage of variance explained by the second component [135]. Also, the following heart rate variability frequency parameters were included: the power in the high frequency (0.15-0.4 Hz, HF) and low frequency (0.04-0.15 Hz, LF) bands, their normalized versions, and the LF/HF. Finally, the PRSA downward curve of the RR-intervals was computed with a window length of 10 seconds [10, 130]. From this PRSA curve, the amplitude

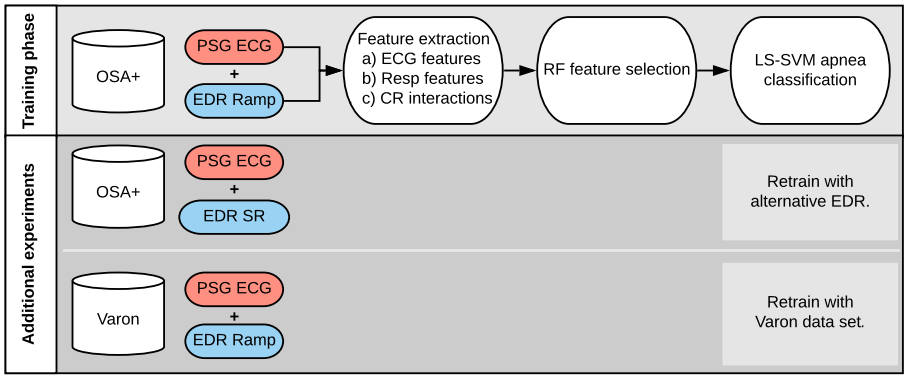


Figure 7.2: Development of the apnea detection classifier. The top row depicts the different steps used to train the classifier. Below, an overview is given of the additional experiments conducted to further analyze the obtained results.

difference between the highest and lowest peaks around the anchor point, overall slope, slope before and after the anchor point, and the accelerating capacity were extracted as features.

Respiratory features: From the EDR signals, 13 features were extracted: the mean, standard deviation and kurtosis; the five PRSA features, described above, extracted from the downward PRSA curve of the respiratory signal with a window length of 10 seconds; and the power of the signal in the HF and LF bands, their normalized versions and LF/HF.

Cardiorespiratory interactions: A large number of features to estimate the cardiorespiratory coupling were investigated. These were computed on both combinations of the ECG and the two EDRs. Cardiorespiratory interactions based on the respiratory bandwidth, information dynamics, bivariate PRSA, time frequency representation and subspace projections were computed [94]. Additionally, the respiratory and RR-interval signals were decomposed in five db4 wavelet levels, from which the respiratory component of the heart rate was extracted using subspace projections as introduced in [135]. The low frequency respiratory variations (WAV_{LF}, 0 - 0.07 Hz) and high frequency respiratory variations (WAV_{HF}, 0.07-0.6 Hz) were extracted.

Feature selection

A total of 117 features were extracted from the different signals, the logarithm was also computed for all of them and considered as feature. In order to find the most relevant features for the detection of sleep apnea, feature selection was performed on the OSA+ set. The RF-based feature selector, as explained in Chapter 5, Section 5.3, was applied.

Classification

Once the relevant features were discovered, a classifier was trained to predict whether or not a 1-minute segment contained an apnea. A FS LS-SVM classifier with RBF kernel was trained based on the OSA+ data set. The 6000 most representative samples were selected for training based on the entropy criterion [123]. The performance of the model was measured using the AUC, the accuracy, sensitivity and specificity. The optimal cut-off point on the ROC was selected based on the optimization of the multiplication of Se and Sp.

The performance results were compared between the two models using the different EDRs. For the Varon set, only EDR Ramp was considered, for which the performance was compared against the results from [135]. The apnea scoring

rules have, however, changed since the Varon data set was recorded, and the patient demographics differ between the Varon and OSA+ data set. Therefore, a classifier trained on the latter data set might not generalize well on the former. In order to ensure a fair comparison to the results obtained in [135], the classifier was retrained on a subset of the Varon set.

7.3 Results

7.3.1 Extracted features

Seven features were identified to be the most relevant ones using the OSA+ data set. These include two ECG features (the log standard deviation of the RR-intervals and $\mathrm{QRS}_{\mathrm{morph}}$), two respiratory features (Log standard deviation and log kurtosis of the respiration) and three cardiorespiratory interaction features ($\mathrm{WAV}_{\mathrm{LF}}$, log $\mathrm{WAV}_{\mathrm{HF}}$ and the log 3-dB bandwidth of the respiratory power spectral density estimate (log PSD BW resp)). These features, extracted using the Ramp EDR from the Imec data set, are plotted in Figure 7.3 for normal (N) and apneic segments (A). When the SR EDR is used, very similar plots were obtained.

7.3.2 Apnea detection

Using the seven extracted features, two classifiers were trained each using one of both EDR estimates. The performance results on the test set of the OSA+ and Imec data sets can be found in Table 7.2. The performances on the Imec data set are slightly better than on the OSA+ test set. On the Imec set, AUCs of 76.3 % (Ramp) and 75.3 % (SR) were obtained. The use of the SR EDR did

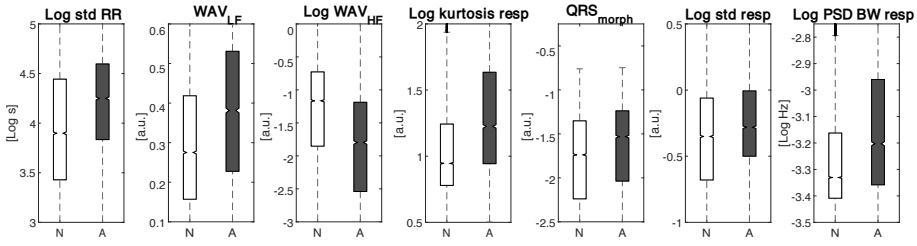


Figure 7.3: Selected features for normal (N) and apneic (A) segments using the Ramp EDR on the Imec data set.

Table 7.2: Apneic segment classification performance based on a combination of ECG and EDR estimates.

	OSA+		Imec	
	Ramp	SR	Ramp	SR
AUC [%]	73.5	73.8	76.3	75.3
Acc [%]	68.8	70.8	68.9	69.8
Se [%]	63.9	57.8	67.0	65.1
Sp [%]	70.8	67.2	69.8	72.0

Table 7.3: Performance comparison to the study by Varon et al. [135].

	Trained on OSA+	Retrained on Varon	Original study [135]
AUC [%]	75.9	85.2	90.0
Acc [%]	68.7	79.9	75.9
Se [%]	66.1	64.1	78.8
Sp [%]	69.9	86.0	84.6

not cause an improvement in performance compared to using the Ramp EDR. For the Varon data set, a similar AUC of 75.9 % was obtained using the EDR Ramp, as can be seen in the first column of Table 7.3.

Retraining the model with the Varon data set

Retraining the classifier using a subset of the Varon data set caused an increase in AUC performance of 10 % to 85.2% for the test portion of this data set, as can be seen in Table 7.3.

7.4 Discussion

7.4.1 Selected features

The seven selected features are very comparable to those selected in [135]. Only the kurtosis of the respiration and the PSD BW resp feature were not considered

by [135], but the serial correlation coefficient was included instead. A lot of different methodologies were tested to estimate the cardiorespiratory interactions, of which the wavelet based subspace projection features were selected in the final set. This might be due to the fact that the wavelet decomposition acts as a filter which smooths out estimation errors in the EDR, while the other cardiorespiratory features take the unfiltered EDR signals as input.

7.4.2 Apnea detection

The proposed algorithm obtained an AUC of 76.3 % when using the Ramp EDR. This performance might seem rather low compared to studies from the literature, in which AUCs of 89-94 % are reported [86]. However, most of the studies use the Physionet Apnea-ECG data set. This data set is rather easy to classify, since the control subjects are volunteers instead of patients with sleep related complaints [104]. Therefore, only a direct comparison to the algorithm by Varon et al. [135] was included, since this clinical data set was available for this study. The results of this comparison using the Ramp EDR can be found in Table 7.3. Applying the pre-trained model developed in this thesis to the Varon data set, resulted in similar performances as in the other two data sets. However, the AUC of 75.9 % is much lower than the AUC of 90.0 % in [135]. This performance drop might be explained by the difference in apnea scoring rules between the OSA+ and Varon data sets, but also by a difference in patient demographics. In Table 2.1, it can be observed that the Varon set contains more men, of which a larger percentage has moderate to severe sleep apnea resulting in a larger average AHI and a smaller range of AHI values. When analyzing the different apnea subtypes, it was discovered that in this data set, of all segments labeled as apneic, containing apneas or hypopneas, 22 % contained obstructive apneas. These are in general the most severe events, and the easiest to automatically detect. Whereas for the OSA+ and Imec data sets this was, respectively, only 10.9 and 11.8 %. When the model was retrained on the Varon data set, the AUC on the Varon test subset increased to 85.2 %. This is still slightly lower than the results in [135]. This drop is due to the lowered sensitivity, which might be caused by the change in feature set since this one was optimized for the OSA+ data set.

Another difficulty with using a clinical data set is the fact that OSA patients often suffer from cardiovascular comorbidities [19]. These comorbidities will distort the RR-interval rhythm and the cardiorespiratory interactions, and might thus distort the extraction of the proposed features. In our study, the presence of CV comorbidities was not considered as an exclusion criterion, since otherwise more than half of the patients coming into the sleep lab for diagnosis would have been excluded. For the OSA+ set, the presence of CV comorbidities

was recorded as explained in Chapter 2. Of those patients, 151 had a CV comorbidity at the time of the PSG, 46 were controls and for the other patients the CV data was incomplete. When the apnea detection performance using the Ramp EDR was compared between them, a slight drop in AUC to 73.1 % was observed for the CV group, compared to an AUC of 75.3 % for the controls. Further investigation into the differences between those two groups also revealed an increased correlation between changes in RR-intervals and the EDR time series, measured using the PRSA parameters, for the CV comorbidity group during normal sleep. This increased correlation was not present when the EDR time series was substituted by the TB. This might indicate that in these patients, the EDR methods do not capture the true cardiorespiratory interactions, but rather just the changes in heart rate, which should be studied further.

In this study, the SR EDR estimate was proposed as an alternative for the Ramp EDR. Since from the comparison in [132], it was concluded that this was one of the most robust EDR methods, we hypothesized that using this alternative EDR an improved performance could be reached. Our results, however, show almost no difference between both EDR methods.

7.5 Conclusion

This chapter presented an algorithm for the automatic detection of sleep apnea based on single lead ECG. A set of ECG, respiratory and cardiorespiratory parameters were extracted from the ECG signal and two EDR estimates. The most relevant features for the detection of apneic events were then selected and a FS LS-SVM classifier was trained. This classifier obtained comparable results on the Varon data set as the paper by Varon et al. [135].

The algorithm was developed such that it also can be applied on capacitively-coupled ECG signals combined with unobtrusive respiratory measurements. In Chapter 9, the algorithm will be applied on a combination of PSG ECG and respiratory effort signals of the thoracic belt. Furthermore, in Chapter 10, the algorithm will be tested on a combination of the non-contact ccECG and ccBioz signals.

Chapter 8

PPG-based sleep apnea detection

This chapter discusses the use of pulse photoplethysmography for the detection of sleep apnea. PPG-time series, known to be modulated by both respiration and the autonomous nervous system, were derived. Moreover, newly defined features based on the point-process model of human heart beat interval dynamics to estimate the autonomic nervous system activity, based on the instantaneous powers in the high and low frequency band of the pulse rate, were introduced. For all these time series, features were extracted over one-minute windows. The most relevant features were selected and a classifier was trained to predict whether the one-minute segment was apneic or not.

Partial results of this algorithm were presented at Computing-in-Cardiology 2018: ***Deviaene, M.****, Lázaro, J., Huysmans, D., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Sleep apnea detection using pulse photoplethysmography. In 2018 Computing in Cardiology Conference (CinC) (2018), vol. 45, IEEE, pp. 1–4.*

8.1 Materials

The PPG-based OSA detection algorithm was developed on a subset of the UZ Leuven OSA+ data set, including 102 recordings. From these, a training set of 71 recordings was selected as the centers of k-medoids clustering on the patient characteristics. The remaining 31 recordings were used as independent test

set. These sets will be referred to as, OSA+$_{\text{train}}$ and OSA+$_{\text{test1}}$. The results presented in [39], were based on these two sets. However in the meantime, the OSA+ data set was extended with 116 recordings. In this chapter, the algorithm will also be tested on these additional recordings, which will be called the OSA+$_{\text{test2}}$ set. Moreover, the Imec data set was also added as test set.

From the PSG recordings, the PPG signals, sampled at 500 Hz, were extracted. The data was split into 1-minute segments, that were labeled apneic if they contained at least 10 seconds of apneic events, which corresponds to the minimum duration to be annotated [17].

8.2 Methods

8.2.1 PPG pre-processing and feature extraction

The PPG signals were preprocessed by applying a low-pass filter with a cut-off frequency of 35 Hz [78]. Next, the PPG peaks were detected by applying a linear filtering transformation to emphasize the peak slopes, followed by an adaptive thresholding [79]. From each detected peak, several points of interest were extracted which will be used to compute the PPG time series, see Figure 8.1. The peak apex point (n_A) was extracted, as well as its basal (n_B) and mid-point (n_M). Moreover, the first derivative of the signal was used to compute the peak onset (n_O), end point (n_E), and the slope end point (nEs) [78].

From these peak points 6 PPG time series were computed. The pulse rate variability (PRV) was computed as the inverse of the time between consecutive n_M's; the pulse amplitude variability (PAV) as the amplitude difference between n_B and n_A; and the pulse width variability (PWV) as the time between n_E and

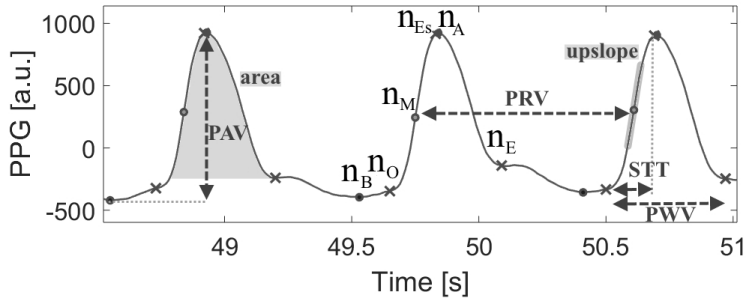


Figure 8.1: Overview of the PPG peak point delineation and feature extraction.

n_O. The slope transit time (STT) was defined as the time in between n_{Es} and n_O. Additionally, the maximal pulse upslope and the area under the peak were computed, as can be seen on Figure 1.

The autonomic nervous system can be assessed by studying the low frequency (0.04-0.15 Hz, LF) and high frequency power (0.15-0.4 Hz, HF) of the PRV. During apneas, however, there is a close succession of parasympathetic and sympathetic changes [19]. Therefore, it is interesting to study these parameters on short time intervals. To do so, the point-process model of human heart beat interval dynamics developed by Barbieri et al. [9] was applied to the PPG peak mid-points n_M, as surrogate of the R-peaks in the ECG. This model uses a history-dependent inverse Gaussian distribution to model the RR-interval time series and provides an instantaneous estimate of the heart rate LF power, the HF power and the LF/HF index [9].

The 9 computed PPG-time series were post-processed using a median absolute deviation (MAD)-based outlier detection rule [79]. Samples were removed, if their feature values were outside the range of [med_y - 5 std_y, med_y + 5 std_y], with med_y and std_y, respectively, the median and standard deviation of the last 50 samples. Outlier segments longer than five times the average heartrate interval were removed, those shorter were interpolated. After outlier detection, the time series were all resampled to 4 Hz and one-minute averaged features were computed. The minimum, maximum and mean value, the standard deviation and the gradient during the one-minute segment were extracted from each of the time series. This resulted in a total of 45 PPG features.

8.2.2 Feature Selection

From the 45 PPG features, the most discriminative ones to separate apneic minutes from normal breathing minutes were selected. A similar feature selection workflow as in the SpO_2-based algorithm of Chapter 6 was applied: first highly correlated features were removed, then the minimal redundancy maximal relevance algorithm [103] was used to further reduce the feature set to 10 features. Finally, any remaining redundant features were removed using a backwards wrapper combined with the classifier explained below.

8.2.3 Classification

Using the selected features, an FS LS-SVM using an RBF kernel was built to classify segments as apneic or normal breathing [123]. In order to balance the training set, an equal number of segments from the apnea and normal class

were included per training patient. All segments from the smallest class were included, the segments from the largest class were selected as the centers of k-medoids clustering on the feature set. The LS-SVM hyper parameters were optimized using 10-fold cross validation on this training subset.

8.3 Results

The PPG feature selection resulted in a subset of 6 features: the minimum of the PAV, standard deviation and mean of the PRV, standard deviation of the PWV and the minimum and mean value of the LF/HF power of the pulse rate. Boxplots of these features for the OSA+ data set are depicted in Figure 8.2, comparing the normal and apneic segments. The obtained classification performances on the test sets are presented in Table 8.1.

8.4 Discussion

When observing the selected PPG features in Figure 8.2, a decrease in PAV is seen during apneic minutes. This decrease has been linked to vasoconstriction

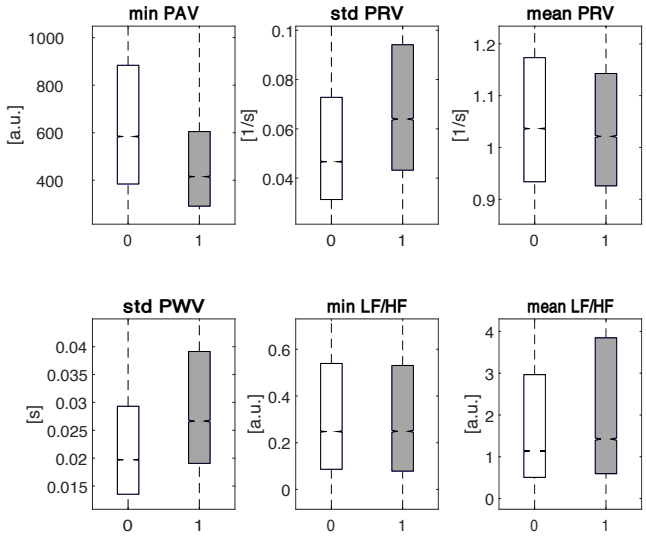


Figure 8.2: Selected PPG features for normal breathing (0) and apneic segments (1) on the OSA+ data set.

Table 8.1: Classification performance of the PPG apnea detection algorithm.

	Acc	Se	Sp	AUC	PPV
OSA+$_{\text{test1}}$	68.7	69.1	68.6	74.8	42.4
OSA+$_{\text{test2}}$	62.8	68.3	60.3	69.8	43.7
Imec	63.5	66.2	62.2	69.2	45.0

and arousals due to apneas [79]. Moreover, an increased standard deviation of the PRV is observed, indicating autonomic nervous system activations and the increased LF/HF power points towards an increased sympathetic activation.

Using the six PPG features, the PPG-based apnea classifier was trained. The PPG classification performances, given in Table 8.2, are lower than in previous studies from the literature. We achieved an accuracy of 68.7 % on the OSA+$_{\text{test1}}$ set, whereas in [78], an accuracy of 72.66 % was obtained on the same data. In this study, however, only obstructive apneas were considered. Our study considers all types of apneic events, including central events and also hypopneas which only have a partial obstruction of the airway, and are thus more difficult to detect. Only 11.9 % of the apneic minutes in this study included obstructive apneas. If only these event minutes are taken into account, the accuracy, sensitivity, specificity and AUC of PPG-based detection on the OSA+$_{\text{test1}}$ set rise to, respectively, 76.2 %, 67.9 %, 76.5 % and 78.0 %. The sensitivity and AUC further increased when central and mixed apneas were added, probably due to the increase in available training data. An accuracy of 74.8 % was then achieved with a sensitivity, specificity and AUC of, respectively, 73.8 %, 74.9 % and 80.4 % on the OSA+$_{\text{test1}}$ set.

A lower performance was obtained on the OSA+$_{\text{test2}}$ and Imec data sets compared to the OSA+$_{\text{test1}}$ set. The AUC dropped from 74.8 % to, respectively, 69.8 and 69.2 %. Moreover, the performance parameters vary a lot between patients. If a patient has only few apneic events, they are often more isolated, have a less severe autonomic response and are therefore more difficult to detect. When apneas are grouped together, they are more easily detected and thus higher sensitivities are obtained. Another observation that can be made is the low positive predictive value of 42.4-45.0 %. Previous studies also encountered the same problem. A lot of sympathetic activations are detected, which are not linked to annotated apneic events. These could be due to subclinical apneas which are too short to be scored according to the AASM rules, or other causes such as movement of the patient. In [78], it was investigated whether these false positives occurred more often during REM sleep, which was not the case. In this study the patients of the OSA+$_{\text{test1}}$ set were on average in REM sleep

for 14 % of the recordings. During this time, 20 % of all apneic minutes, and 20 % of all false positives occurred. There is thus no excessive amount of false positive detections during REM sleep. Therefore, REM sleep autonomous nervous system variations cannot account for the low PPV.

The obtained results might still be improved if extra PPG features were considered, investigating for example the dynamics and nonlinear properties of the PPG signal. Moreover, the full night recording was processed, including periods where the patient lies awake during the night. Implementing some PPG-based sleep staging could avoid false positives during wake, which now represent 22.4 % of the false positives in the $\text{OSA}+_{\text{test1}}$ set.

The SpO_2-based apnea detector from Chapter 6 clearly outperforms the presented PPG-based classifier. The SpO_2 signal is, however, not always available in wearable PPG measurement systems, such as in popular smartwatches. The ECG-based detector from Chapter 7 also performs slightly better. This might be due to the fact that the ECG signal is less affected by noise, and the fact that the respiration has a more pronounced effect on the ECG, making it easier to extract respiratory changes from the ECG than the PPG signals.

8.5 Conclusion

This chapter investigated the use of PPG derived features for the detection of apneic minutes in overnight recordings. The PPG-based results are promising, but the positive predictive value should still be improved to be useful in OSA screening applications.

In Chapter 9, the added value of the proposed PPG features to the SpO_2-based OSA detection will be investigated. Furthermore, the use of this PPG-based OSA detection algorithm will be tested on wearable Byteflies PPG data in Chapter 10.

Chapter 9

Multimodal sleep apnea detection

This chapter investigates whether the apnea detection performance can be improved by combining different modalities. First, the combination of ECG and respiratory effort signals is presented in Section 9.1. The thoracic belt will be introduced in the ECG-based apnea detection algorithm, developed in Chapter 7, as a replacement for the EDR estimate. The classification performance will be compared with the original model in order to quantify the added value of the TB signal. The obtained results were included in: ***Deviaene, M.****, Castro, I. D., Borzée, P., Patel, A., Torfs, T., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Capacitively-coupled ECG and respiration for the detection of sleep apnea, submitted to Physiological Measurement.*

Next, the added value of adding PPG and ECG features to the SpO_2-based classifier will be discussed in Section 9.2. Preliminary results of this analysis were published in ***Deviaene, M.****, Lázaro, J., Huysmans, D., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Sleep apnea detection using pulse photoplethysmography. In 2018 Computing in Cardiology Conference (CinC) (2018), vol. 45, IEEE, pp. 1–4.*

Additionally, the use of a hidden Markov model, based on SpO_2 and PPG features was investigated in Section 9.3. The chapter will be concluded with Section 9.4 on the effect that the inclusion of automatic sleep scoring has on the OSA screening performances.

9.1 OSA detection based on ECG and respiratory effort

9.1.1 Methods

The same workflow was followed as in the original algorithm, described in Chapter 7. First, the ECG and TB signals were preprocessed. Similarly as the ECG signals, the TB signals were upsampled from 500 Hz to 512 Hz. The PSG recording system already prefiltered the TB signal internally, and no additional prefiltering was required. Neither was any artefact detection implemented, as it was assumed that movement artefacts would overlap with the ones detected in the ECG signal. The robust standardization described for the EDR in Chapter 7 was applied to the upsampled TB signals and the signals were divided into 1-minute windows.

Next, the relevant subset of seven features was extracted for each 1-minute window. The log standard deviation of the RR-intervals and $\text{QRS}_{\text{morph}}$ feature were extracted from the ECG. The log standard deviation and log kurtosis were computed on the TB signal. Additionally, the three cardiorespiratory interaction features (WAV_{LF}, log WAV_{HF} and log PSD BW resp) were extracted from the combination of the ECG and TB signals.

Using these features, the classifier was retrained on the OSA+ train set and tested on the OSA+ test set and the Imec data set. As such, the added value of the TB over the EDR estimate could be quantified.

9.1.2 Results

In Figure 9.1, the extracted respiratory and CR interaction features are plotted for the Imec data set. When compared with the features extracted using the EDR in Figure 7.3, it can be observed that the log kurtosis shows a much better separation between normal and apneic segments when the feature is extracted from the TB than from the EDR estimates. The other features show no clear visual differences.

The performance results for the retrained model, using the combination of the ECG signals with the TB, on the test set of the OSA+ data set and the Imec data can be found in Table 9.1. When using the TB, an AUC of 80 % is obtained on the Imec data set, whereas the EDR combinations only obtained AUCs of 76.3 (Ramp) and 75.3 % (SR).

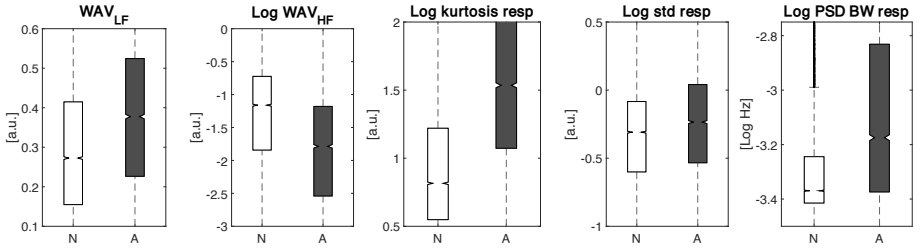


Figure 9.1: Respiratory and CR interaction features for normal (N) and apneic (A) segments using the TB on the Imec data set.

Table 9.1: Apneic segment classification performance on the retrained classifier based on the combination of ECG and TB signals.

	OSA+	Imec
AUC [%]	79.5	80.0
Acc [%]	74.2	72.9
Se [%]	67.1	69.5
Sp [%]	77.1	74.5

9.2 Addition of PPG and ECG features to SpO_2-based apnea detection

The SpO_2-based model has shown the best unimodal OSA detection performance. It, however, has as a main disadvantage that only events accompanied with an oxygen desaturation can be detected. Apneas without an oxygen desaturation and hypopneas linked only to an arousal will be missed. Other modalities should be added to this classifier to detect these missed events. Since the pulse oximeter, used to measure the SpO_2 also records the pulse photoplethysmography signal, this signal seems to be the first choice to improve the SpO_2-based detection. From the PPG signal, parameters linked to the autonomic arousals and respiration can be extracted. Also, the added value of the ECG features will be tested, as slightly better unimodal classification performances could be reached using these features compared to the PPG ones.

This approach makes use of the SpO_2-based model developed in Chapter 6, the PPG-based model from Chapter 8 and the ECG-based model of Chapter 7.

9.2.1 Materials and methods

Data sets

For the training and testing of the combined classifier, the same data sets as discussed in Chapter 8 were used. 71 subjects of the UZ Leuven OSA+ test set were used for training, while the rest of the OSA+ data set and the Imec data set were used for testing. All 147 OSA+ test subjects were merged in one test set, denoted $\text{OSA+}_{\text{test}}$. The full-night recordings were split into 1-minute windows, which were labeled apneic when they overlapped for at least 10 seconds with an apnea or hypopnea.

SpO_2 processing

For the SpO_2 processing and feature extraction, the workflow described in Chapter 6 was used. Oxygen desaturations were detected and six features were extracted from each of them: the SpO_2 variance, the amplitude of desaturation, the length of the resaturation period, the amplitude of the upward and downward PRSA [10] curves, and the timing of the first peak in the autocorrelation of the SpO_2 signal. These features were then used to classify the desaturations as apneic or not.

In this study, 1-minute segments were analyzed. Each desaturation was linked to a segment, taking into account the delay of about 30 seconds between the start of the apnea and the start of the desaturation [41]. Therefore, desaturations were linked to the segment if they start earlier than 30 s after the end of the segment and stop later than 30 s after the start of the segment. The segment got the SpO_2 feature values of the linked desaturation with the highest probability to be apneic. When no desaturation was linked to a segment, the SpO_2 features were set to NaN.

PPG processing

The PPG pre-processing and feature extraction explained in Chapter 8 was applied. Six PPG features were extracted: the minimum of the PAV, the standard deviation and mean of the PRV, the standard deviation of the PWV and the minimum and mean value of the LF/HF power of the pulse rate.

ECG processing

Seven features were extracted from the ECG signals in combination with the R_{amp} and SR EDR estimates as explained in Chapter 7, as well as from the combination of the ECG and the TB signal, as explained in Section 9.1.

Classifier training

A fixed-size LS-SVM classifier with an RBF kernel was trained on the combination of the SpO_2 and PPG feature sets, as well as combinations of the SpO_2 feature set with the different ECG feature sets. The training set was balanced as discussed in Chapter 8 and the classifier's hyperparameters were optimized using 10-fold cross-validation. In order to assess the added value of the ECG and PPG features for 1-minute based classification, the classifier was also retrained only including the SpO_2 features. The 1-minute classifiers only based on the ECG and PPG features were already presented in, respectively, Chapters 7 and 8.

9.2.2 Results

The classification performances for the combined models and the reference 1-minute SpO_2 classifier are given in Table 9.2. The SpO_2 reference model classifies the 1-minute segments with an AUC of 89.5 % for the OSA+ test set and 92.5 % for the Imec set. This performance is higher than the AUC of 86.0 %, obtained using the event-based classification model, as shown in Table 6.9. An increase in sensitivity of 6 % was obtained compared to these results, this is due to the fact that multiple apneic events can occur within a 1-minute window. The event-based model might have difficulties to detect every single one of these apneas, while in the 1-minute model the cumulative effect of successive apneas might improve the detection.

When PPG and ECG features were added to the SpO_2-based model, only slight increases in performance were observed. The ECG-based model using the SR EDR obtained the best performance with an increase in AUC of 0.5 % to, respectively, 90.0 and 93.0 %. Using the TB respiratory effort to estimate the respiratory features and CR interactions, did not improve the performance. The PPG and ECG-based models have shown promising apnea detection results in the previous chapters, but when the SpO_2 signal is available, the added value of these modalities is limited. One of the main objectives was to improve the sensitivity of the SpO_2-based apnea detector by adding the PPG or ECG features. Since extra apneas, which were not linked to clear oxygen desaturations, could

then be detected. The low PPV of the PPG and ECG-based classifiers, however, made it impossible to improve the sensitivity without greatly decreasing the specificity of the classifier. Moreover, applying a 1-minute window approach instead of the event-based classifier, already improved the sensitivity of the SpO_2-based model.

This straightforward combination of features from different signal modalities did not improve the SpO_2 classification performance. However, these results might be improved if the feature selection was rerun on the complete multimodal feature set. Additionally, a hierarchical classification structure could also be investigated. The SpO_2-based algorithm could be used as a first step, after which negative segments are classified using ECG or PPG features. Moreover, PPG and ECG signals have been studied to perform automated sleep staging [52]. Adding these modalities in order to remove false positives during wake and to get a better estimate of the sleep time might thus be effective. In the next section, the use of a hidden Markov model (HMM) including a wake state for the detection of sleep apnea using SpO_2 and PPG signals will be discussed. The final section of this chapter will investigate the AHI estimation and OSA screening capacities of the different signal modality combinations, and which influence the removal of wake regions might have on these results.

9.3 Apnea detection using HMM based on SpO_2 and PPG features

The PPG signal has been used to develop algorithms for automatic sleep staging [52], it can therefore be used to detect wake periods during the night. This

Table 9.2: Classification performance of the combined apnea detection algorithms including SpO_2, PPG and ECG parameters. The reference 1-minute SpO_2-based classifier is presented in the bottom.

	OSA+$_{test}$					Imec				
	Acc	Se	Sp	AUC	PPV	Acc	Se	Sp	AUC	PPV
PPG + SpO_2	83.0	80.2	84.2	89.8	**68.9**	85.0	88.2	83.5	92.6	71.4
ECG R_{amp} + SpO_2	82.2	77.3	84.0	87.0	64.4	83.4	84.1	83.1	89.4	69.1
ECG SR + SpO_2	**83.5**	80.2	**84.7**	**90.0**	66.2	**85.7**	**88.9**	**84.2**	**93.0**	**71.6**
ECG TB + SpO_2	81.5	74.9	84.0	85.7	63.7	82.9	80.9	83.9	88.7	69.2
SpO_2	82.7	**80.3**	83.7	89.5	66.6	85.2	87.9	84.1	92.5	70.4

can improve apnea detection algorithms, as often false positive predictions are made during wake periods. Moreover, the effect of apneic events does not always occur at the same moment for the different signal modalities. Oxygen desaturations, for example usually have a delay of about 30 seconds with respect to the start of the apneic events. Whereas, in the heart rate, at the beginning of the apnea, bradycardia is observed, while tachycardia occurs at the end due to the autonomic arousal. It might therefore be useful to separate the immediate effects of apneas, which should coincide with the apnea annotation based on the respiration and the delayed apneic responses. By modeling the transition from normal sleep, to apnea and apneic response, a more precise apnea detection algorithm can be formed.

In this study, it was investigated whether the use of a hidden Markov model with four states: wake, normal sleep, apnea and apneic response, could be useful for the detection of sleep apnea. By implementing the state transitions and time dependencies of the apneic responses, a better combination of the SpO_2 and PPG features might be possible as the delay between the oxygen desaturation and heart rate responses can be taken into account.

9.3.1 Hidden Markov model for sleep apnea detection

In a hidden Markov model, one tries to model the unobserved underlying state sequence based on some observed parameters [109]. In this application, the observed parameters consist of changes in the SpO_2 and PPG signals. Whereas, the state sequence corresponds to the transition from normal sleep to apneas. Four states were defined: wake, normal sleep, apnea and apnea response. In order to model a HMM for these states based on the SpO_2 and PPG features, different probabilities needed be estimated:

- The initial state distribution.
- The transition probability, which gives for each state the probability to go to all other states.
- The emission probability, which consists of the probability density function of observing a certain set of SpO_2 and PPG features given the current state.

The initial state distribution and the transition probabilities were estimated from state labels of the training data, whereas the emission probabilities were obtained by fitting a Gaussian mixture model for each of the states. After completely defining the HMM, the Viterbi algorithm [109] was used to find the most probable state path based on the emitted SpO_2 and PPG features.

9.3.2 Materials and methods

Data set and signal segmentation

A subset of 102 subjects of the UZ Leuven OSA+ data set was used for this study. 71 recordings were used for training the model, whereas the other 31 were used as independent test set. The recordings were split into 10 second windows, the hidden state should be estimated for each of these windows. The state labels were defined as following:

1. **Apnea**: A 20 second window starting from the PSG apnea start annotation. When the apneic event was shorter than 20 seconds, the apnea state ended at the PSG apnea end annotation.
2. **Apnea response**: Window starting from the end of the Apnea state until the PSG apnea end annotation, with a minimum duration of 20 seconds.
3. **Wake**: Wake segments as annotated in the hypnogram.
4. **Normal sleep**: If none of the above states are valid.

The 10 second windows received the apnea state label if they overlapped 5 seconds with the defined apnea state. When this was not the case, but an overlap of 5 seconds with the apnea response state was observed, the window was assigned as apnea response. Remaining segments that overlapped 5 seconds with wake periods, were labeled as wake. All remaining segments were considered as normal sleep.

SpO_2 and PPG features

PPG and SpO_2 features were calculated over each 10 second time window. The PPG signals were processed as described in Chapter 8 and the described nine PPG time-series were extracted. The mean values of these time series, and the difference with the previous window were considered as features. From the SpO_2 signals, all desaturations were detected as described in Chapter 6 and the six selected features were extracted from each desaturation. Each window received the SpO_2 features, of the desaturation which overlapped with the window. Moreover, also the mean SpO_2 value within the 10 second window, and the difference of this mean with the mean of the previous window were considered.

Model optimization

Different model optimization steps were needed. First, the optimal sequence length needed to be defined. Choosing a longer sequence length can lead to a more precise estimation of the state sequences, but will also result in a longer computation time. Moreover, sequences which include artefacts had to be removed. Therefore, a sequence length of 10 minutes was chosen to minimize the amount of artefact sequences.

Next, feature selection was performed using the RF-based feature selector explained in Chapter 5. 13 features were retained, including 4 SpO_2 features (the amplitude of desaturation, the SpO_2 PRSA upward and downward amplitude difference and the mean SpO_2 value), and 9 PPG features (the mean PRV, PWV, STT, area under the peak, LF, HF and LF/HF, and the PRV and slope differences).

Using the selected features, a Gaussian mixture model (GMM) was fitted for each state. 10-fold cross validation on the final classification performance was used to optimize the model based on the obtained κ value. The number of Gaussian mixtures was optimized, which led to the selection of 7 Gaussian functions.

9.3.3 Results

An overview of the HMM states with their transition probabilities can be seen in Figure 9.2. When a subject is in the wake or normal sleep state, there is a large probability of over 95 % that in the next 10 second window the subject will remain in that state. When the subject is in the apnea state, either the apnea can continue in the next window, or the subject can go into the apnea response state. Due to the state definitions, the sequence will always end up in the apnea response state after the apnea state. From the apnea response state, the subject has a probability of 15.1 % of going to normal sleep. However, in 12.6 % of the times, the subject will go again to the apnea phase. This shows the instability of the breathing of OSA patients. Stable breathing is often not regained, but apneas follow in close succession.

The performance results of the HMM state estimation using the Viterbi algorithm are shown in Table 9.3. The highest κ score was obtained for the normal sleep stage, which had the best detection performance. This might be due to the fact that this state was most present in the data set. When normal sleep was misclassified, it was often mistaken as the wake state. The apnea and apnea response state have low sensitivities of 47 and 41 %, this is partially due

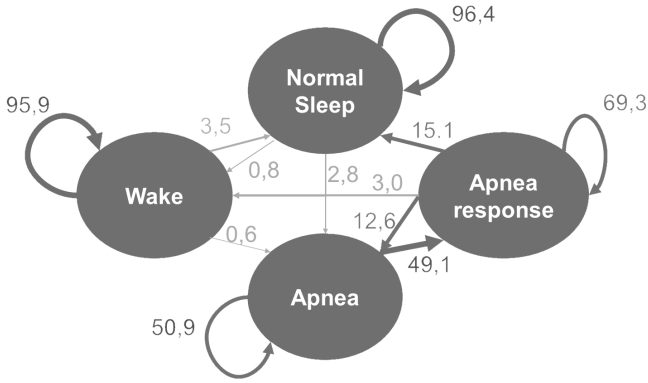


Figure 9.2: Overview of the HMM states and their transition probabilities.

Table 9.3: HMM state estimation performance in %.

	Acc	Se	Sp	PPV	κ
Wake	77	59	81	47	37
Normal Sleep	76	71	82	80	53
Apnea	87	47	93	47	39
Apnea Response	84	41	92	55	44

to the fact that both states are often confused. When these two states were taken together, the detection sensitivity increased to 65 %.

The obtained classification results are promising, however, when the state classification was performed using only a GMM, very similar results were obtained. The HMM sequence information thus did not have any added value in this model.

9.4 Added value of sleep staging for OSA screening

In this section, the AHI estimation and OSA screening performance will be compared for the different unimodal and multimodal 1-minute apnea detection models explained in Section 9.2. Moreover, the added value of implementing sleep staging into the models will be investigated by using the hypnogram

labels available from the PSGs. In the future, these sleep/wake labels could be extracted automatically based on the ECG or PPG signals [52].

9.4.1 Methods

Apnea detection was performed on a 1-minute window base for each signal modality, and signal modality combinations as explained in Section 9.2. Afterwards, the AHI was computed as the number of windows classified as apneic, divided by the number of windows without artefacts, multiplied by 60. A robust linear regression was then performed on the training set of the OSA+ data set in order to obtain a better estimate of the AHI as discussed in Chapter 6. Next, an OSA screening classifier with an AHI threshold of 15 was trained by optimizing the threshold on the estimated AHI, such that the multiplication of Se and Sp was maximized.

This procedure was then repeated with an alternative AHI estimate in which 1-minute windows which were annotated completely as awake were removed. As such, false positive detections during wake were removed, and the number of apneas can be divided by the time asleep, rather than the recording time. The scoring of sleep stages was performed based on 30 second windows, thus wake periods were only removed if 2 consecutive wake annotations were given, as the full 1-minute window had to be scored as wake in order to be removed. As such, short awakenings which might be caused by apneic arousals were kept in the data set and only longer wake periods were removed.

9.4.2 Results

In table 9.4, an overview of the coefficients of determination of the linear regressions between the estimated and true AHI values for the different signal modalities is given for the OSA+ test set and the Imec data set. The R^2 values are given with and without the removal of wake segments. The SpO_2-based AHI estimates have the highest R^2 values, of respectively, 0.836 and 0.934, for the OSA+ and Imec test sets. A small improvement in R^2 values is observed when the AHIs were estimated while removing the wake periods. Including ECG and PPG features to the model did not improve the SpO_2-based R^2 values. The PPG and ECG-based AHI estimates obtain much lower R^2 values between 0.113 and 0.613. The best results were obtained when the ECG signal was combined with the TB respiratory effort signal. Again, only small increases in R^2 values were seen when the wake segments were removed from the analysis. Moreover, better results were obtained on the Imec data set than the OSA+ set.

Figures 9.3 to 9.6 obtain scatter plots of the true AHI values versus the estimated ones for the different unimodal OSA detection models for the Imec data set. As expected from the results from Table 9.4, the best estimations are seen for the SpO_2-based AHI estimation in Figure 9.3, followed by the AHI estimate based on the ECG signal combined with the thoracic belt in Figure 9.6. Only small visual changes can be observed when the AHI was estimated including the removal of wake segments.

Table 9.4: R^2 values for the linear regression between the estimated and true AHI values for the different signal modalities.

	No wake removal		Wake removal	
	OSA+	Imec	OSA+	Imec
SpO_2	0.836	0.934	0.886	0.950
PPG	0.113	0.199	0.157	0.237
ECG SR	0.115	0.431	0.212	0.521
ECG TB	0.318	0.613	0.396	0.667
SpO_2 + PPG	0.815	0.935	0.879	0.959
SpO_2 + ECG SR	0.760	0.937	0.823	0.951
SpO_2 + ECG TB	0.725	0.922	0.794	0.937

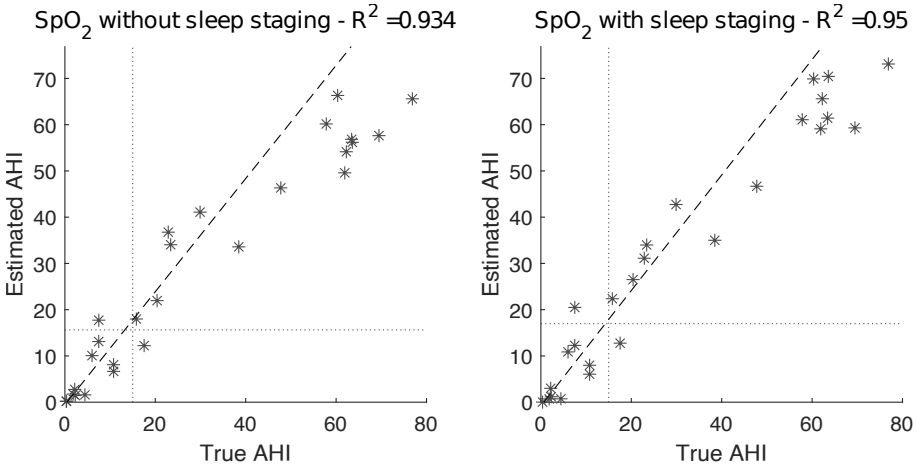


Figure 9.3: Estimated vs true AHI value for SpO_2-based OSA detection with and without including sleep staging on the Imec data set.

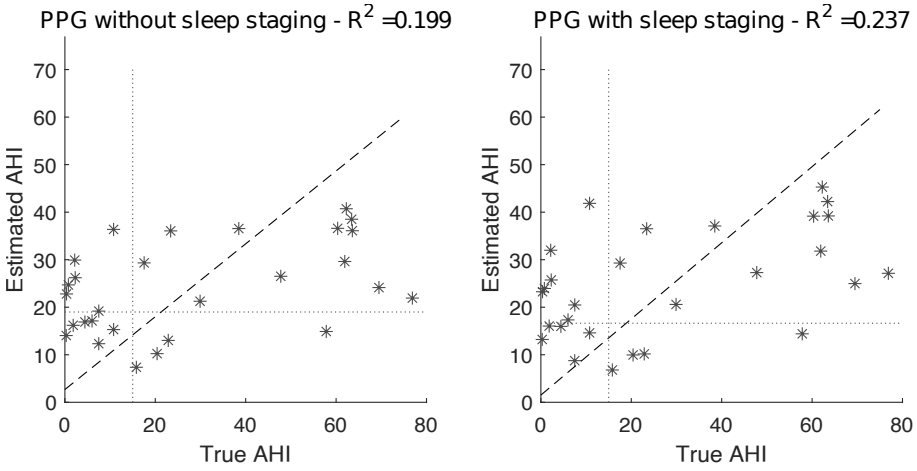


Figure 9.4: Estimated vs true AHI value for PPG-based OSA detection with and without including sleep staging on the Imec data set.

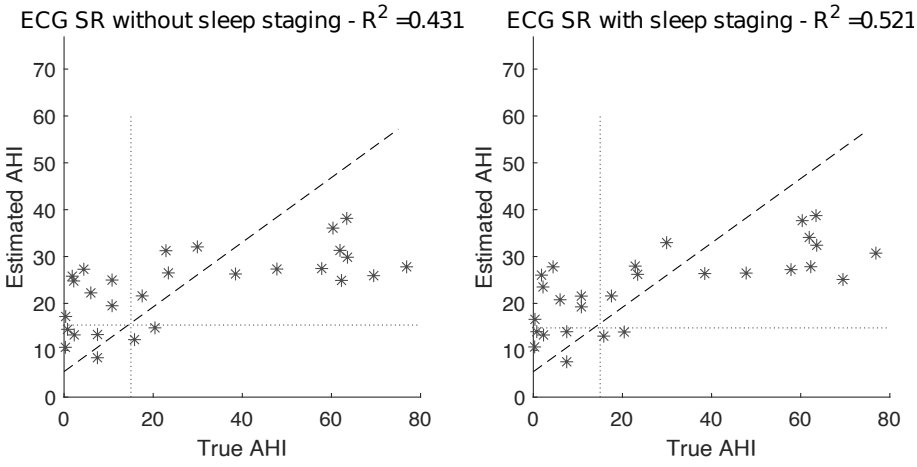


Figure 9.5: Estimated vs true AHI value for ECG SR EDR based OSA detection with and without including sleep staging on the Imec data set.

In Table 9.5, the OSA screening performances with an AHI threshold of 15 for the different signal modalities and combinations with and without removing wake segments is given for the OSA+ test set. The SpO_2-based screening performance is similar to what was obtained in Chapter 6. The screening performance did not improve when wake segments were deleted. Moreover, the

addition of PPG and ECG features did not improve the performance. The PPG and ECG-based models obtain lower screening performances, which are not useful in clinical practice. AUCs of about 63 % were obtained, and the AUC increased to 73 % when the TB was combined with the ECG signal. Again, the removal of wake segments only caused increases in performance of a few percentages. On the Imec data set, slightly better results were obtained, especially for the ECG-based methods. The ECG SR method obtained an AUC

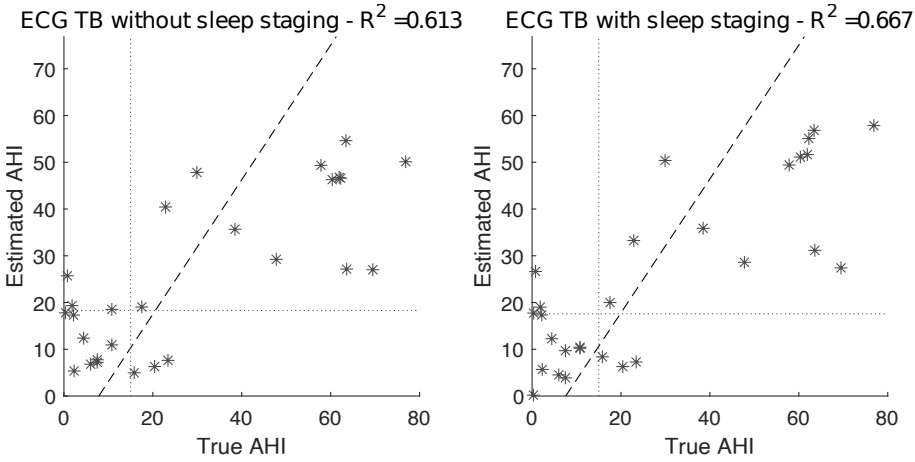


Figure 9.6: Estimated vs true AHI value for ECG TB based OSA detection with and without including sleep staging on the Imec data set.

Table 9.5: OSA screening performances with an AHI threshold of 15 for the different signal modalities with and without removing wake segments for the OSA+ test set.

	No wake removal				Wake removal			
	Acc	Se	Sp	AUC	Acc	Se	Sp	AUC
SpO_2	85.7	88.6	81.4	92.8	85.7	86.4	84.7	92.7
PPG	61.9	73.9	44.1	63.8	66.0	84.1	39.0	65.7
ECG SR	61.2	92.0	15.3	63.2	61.9	94.3	13.6	66.0
ECG TB	65.3	68.2	61.0	73.1	66.0	71.6	57.6	74.0
SpO_2 + PPG	83.7	80.7	88.1	91.9	85.0	84.1	86.4	92.4
SpO_2 + ECG SR	81.6	89.8	69.5	91.2	87.8	81.8	96.6	92.5
SpO_2 + ECG TB	84.4	79.5	91.5	90.0	85.7	85.2	86.4	91.1

of 85.4 % with an accuracy and specificity of, respectively, 67.9 % and 41.7 %. Using the TB an AUC of 83.9% was obtained with an accuracy of 78.6 % and a specificity of 75 %. This improved performance can be explained by the fact that in this data set, less subjects with an AHI close to the threshold of 15 were included. Only results using the SR EDR estimate were presented, but the results using the Ramp EDR were very similar.

9.5 Conclusion

In this chapter, several methods to merge the unimodal OSA detection algorithms to multimodal models were investigated. Replacing the ECG EDR estimates by the thoracic belt respiratory effort signal led to a clear increase in performance. From the comparison of feature values, it seems that this increased performance is mainly caused by a better separation between apneic and normal segments for the kurtosis of the TB. Characteristics of this feature were not completely captured by the EDR estimates, leading to a worse classification performance. Although the addition of the thoracic belt led to an increase in performance, adding this extra sensor also makes the measurement system less unobtrusive. A trade-off should thus be made between the patient comfort and the apnea detection performance.

The SpO_2-based model could not be improved by adding PPG or ECG features in a straightforward way. The cardiovascular features have shown promising results in the unimodal classifiers, but in this straightforward combination they have a limited added value compared to the SpO_2-based features. Further investigation is needed into how these modalities might be combined in a smarter way.

One alternative combination that was investigated was the use of a 4-state hidden Markov model including wake, normal sleep, apnea and apnea response, to model the sequence of apneas during the night. This HMM was trained based on SpO_2 and PPG features extracted from 10 second windows. Promising results were obtained, however, upon further analysis it was shown that a simple Gaussian mixture model obtained similar performances. Adding the state sequence information thus did not have any added value in this model.

The last experiment investigated the AHI estimation and OSA screening performances for the different signal modalities with and without removing wake segments. It was concluded that the SpO_2-based model gives an AHI estimate which can be used in clinical practice to screen for OSA, with an AUC of 92.8 %. Adding the PPG or ECG features did not improve this screening performance, nor did the removal of wake segments. Removing wake segments

slightly increased the R^2 values, due to an improved AHI estimation for a limited number of recordings. Overall, however, this did not improve the screening performance. The PPG and ECG-based models only obtained screening AUCs between 63.8 and 73.1 % which is too low to be useful in practice.

Chapter 10

Wearable sleep apnea detection

In this chapter, the algorithms developed based on PSG signals in Chapter 7 and 8 are tested on wearable measurements. In the first section, the use of the ECG-based algorithm is tested on ccECG signals combined with either the EDR estimate or the ccBioz. The obtained results on the capacitively-coupled signals were included in ***Deviaene, M.****, Castro, I. D., Borzée, P., Patel, A., Torfs, T., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Capacitively-coupled ECG and respiration for the detection of sleep apnea, which was submitted to Physiological Measurement.*

The second part of this chapter will focus on the use of the PPG-based OSA detection algorithm on the wearable PPG signals extracted from the reflective Byteflies PPG sensor dots.

10.1 Capacitively-coupled ECG and respiration for the detection of sleep apnea

Although using only an ECG patch for the detection of sleep apnea is already much more comfortable for the patient than the full PSG, electrode-skin-contact is still required, which makes it not ideal for long-term monitoring. As a consequence, non-contact ECG sensors have been developed. Here, the use of a capacitively-coupled ECG (ccECG) sensor, integrated in the mattress, was used

[26]. The sensor also includes a capacitively-coupled bioimpedance (ccBioz) measurement, which can be used to extract the respiratory signal [26].

The ECG-based apnea detection algorithm, which was introduced in Chapter 7 and further tested on a combination of the ECG and the TB in Chapter 9, will now be tested on these capacitively-coupled signals. The models that were pre-trained on the OSA+ data set, using ECG combined with either the EDR or the TB signal, will be tested on the capacitively-coupled measurements of the Imec data set. This set contains 28 ccECG and 11 ccBioz recordings that were simultaneously recorded with the PSG. As such, the usability and limitations of ccECG and ccBioz measurements for the automatic detection of sleep apnea were explored. To our knowledge, this is the first, study investigating automatic OSA detection based on ccECG and ccBioz recordings.

This study builds upon the work described in [135] and [25]. Here, additional cardiorespiratory measures were investigated, and features extracted from the gold-standard PSG signals and the capacitively-coupled measurements were compared. A set of relevant features were selected with and without taking into account if a feature behaves similar when extracted from the PSG and capacitively-coupled signals. Furthermore, the use of a state-of-the-art EDR signal was evaluated for the first time on ccECG.

10.1.1 Methods

The same workflow as in the original algorithm was applied to process the ccECG and ccBioz signals. First the signals were preprocessed, specific adaptations of this preprocessing to the capacitively-coupled signals will be discussed. Next, features were extracted and the pre-trained models were applied. Additionally, some extra experiments were performed to further analyze the obtained results.

Preprocessing

The signal preprocessing steps were adapted to the capacitively-coupled signals. In Figure 10.1, an overview is given of the different preprocessing steps.

ccECG signal: The ccECG signal was prefiltered similar as the ECG signals, the 50 Hz power line interference was removed using a notch filter and a band pass butterworth filter with cut-off frequencies of 0.67 and 40 Hz was applied. Next, the 1-minute ccECG windows were checked for signal inversion. The ccECG array sensor will select every minute the best two electrodes to record the ccECG [26]. Therefore, signal inversion can occur between segments for the ccECG. Afterwards, artefact detection was performed using the algorithm

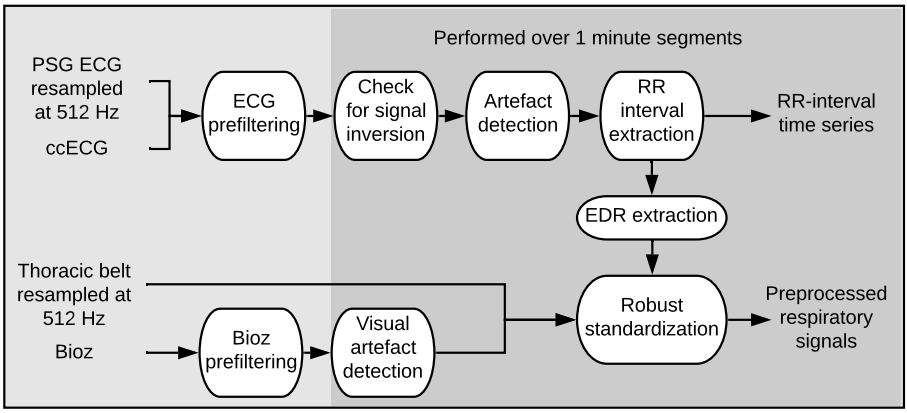


Figure 10.1: Overview of the applied signal preprocessing methods.

proposed by Castro et al. [24]. In this case, only the high quality data segments were included using the high-threshold linear SVM model from [24]. The RR-interval and EDR extraction was finally performed as described in Chapter 7.

ccBioz signal: An extra prefiltering was added for the bioimpedance signals, they were lowpass filtered at 1 Hz. After which, a visual quality check was performed per 1-minute window in order to detect artefacts. Noisy segments were removed. Finally, the robust standardization was applied to the ccBioz signal and the EDRs extracted from the ccECG.

An overview of the available number of clean hours of data per signal modality for the Imec data set can be found in Table 10.1. In Figure 10.2 the ccECG and ccBioz signals are plotted for a 1-minute window which includes an apneic event. The PSG ECG and TB signals were also plotted as reference.

Table 10.1: Overview of the number of hours of clean data for each signal modality of the Imec data set.

Hours of clean data				Total hours of recording	
PSG		Capacitive sensor			
ECG	TB	ccECG	ccBioz	Normal	Apneic
193	254	74	19	172	83

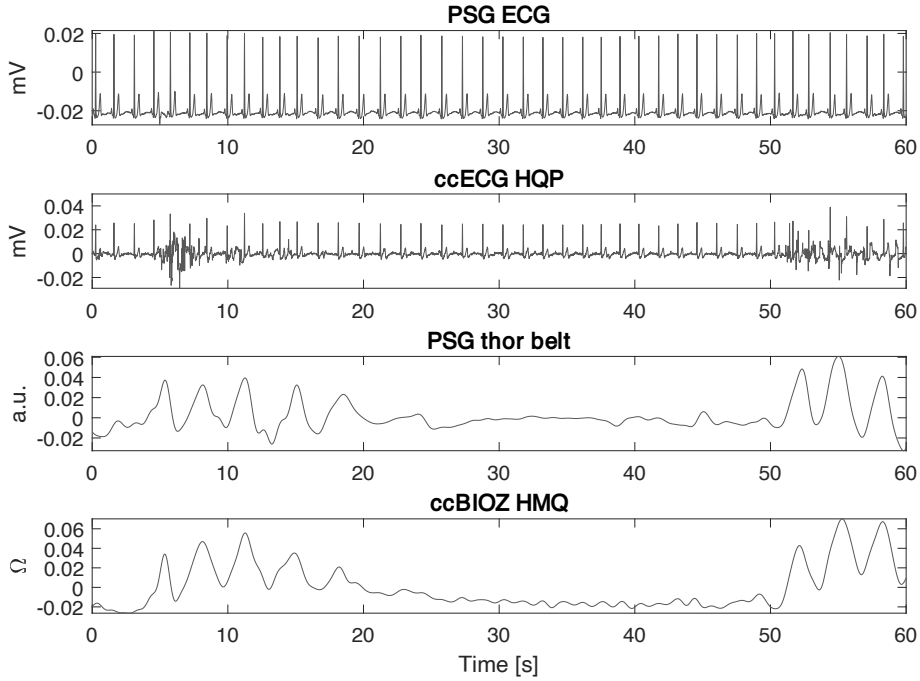


Figure 10.2: Example of preprocessed ccECG and ccBioz signals on a 1-minute segment including an apneic event. The PSG ECG and TB were included as reference signals.

Apnea detection classifier

The classification performance of the pre-trained models on different combinations of ECG and respiratory measurements of the PSG and capacitively-coupled data were compared for the Imec data set, as shown in Figure 10.3. Therefore, the ECG, respiratory and CR interaction features were extracted for all these signal modality combinations.

Additional experiments

In order to further investigate the results, additional experiments were conducted. An overview of these experiments can be found in Figure 10.4.

Comparison of features extracted from PSG and capacitively-coupled measurements: The feature selection was performed on the PSG-based features from the training data set. But, as concluded in [25], not all ECG-

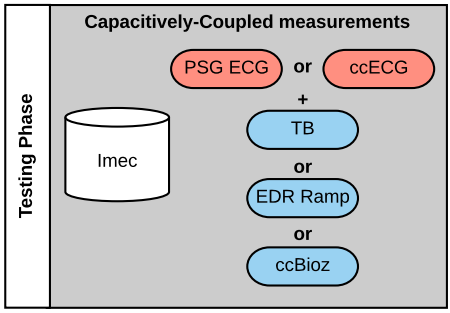


Figure 10.3: Overview of the used signal modality combinations for testing.

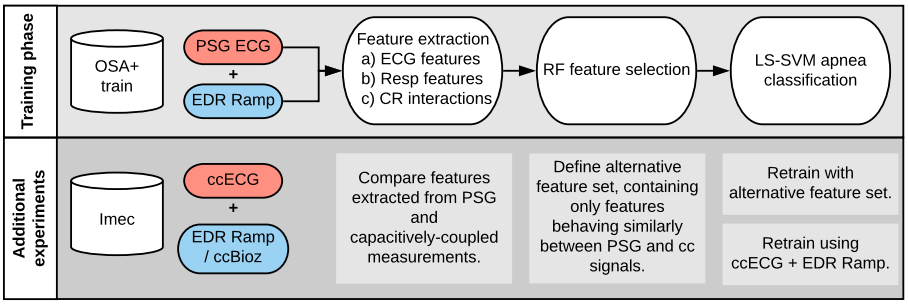


Figure 10.4: Overview of the additional experiments.

based features can be extracted accurately from the ccECG signals. The obtained features might therefore not be the most relevant for the ccECG. In order to investigate the relationship between features extracted from the ccECG and PSG ECG, a linear regression was performed in which the ccECG features were used as regressors and the PSG features as target values. The R-squared values of these regressions were used to evaluate how well the PSG features could be extracted from the capacitively-coupled signals.

An additional feature selection was then performed in which only features which had an R-squared value above a certain threshold were considered. Based on this new feature set, an alternative classification model was retrained using the PSG data of the OSA+ training set.

Retraining the classifier based on capacitively-coupled data: In order to further optimize the classification model towards the capacitively-coupled measurements, the classifier was retrained based on the ccECG and EDR Ramp data of a subset of 20 patients of the Imec set. Unfortunately, not enough ccBioz data was available to be included in this experiment.

10.1.2 Results

Classification features from the original model

In Figures 10.5-10.7, four of the features of the original model are indicated for different combinations of ECG and respiratory measurements for the Imec data set. In Figure 10.5, the QRS_{morph} and log std RR features are compared when extracted from the PSG ECG and ccECG. It is clear that the separation between normal (N) and apneic (A) segments is better for the PSG ECG signals than for the ccECG, especially when using the morphology-related feature, as suggested in [25]. Figure 10.6 depicts the log kurtosis of the different respiratory measures. The TB and ccBioz respiratory signals show a better separation than when EDR signals were used. The selected cardiorespiratory interaction features, on the other hand, show a similar separation for all signal modality combinations, as can be seen for log WAV_{HF} in Figure 10.7.

Classification results using the original model

The classification results on the Imec data for segments with clean capacitive sensor data are presented in Table 10.2. Overall, lower sensitivities can be observed in the reduced data set, even for the PSG-based classification, compared to the results on the full PSG data set, which were presented in Table 7.2. Nevertheless, when the combination of ECG with the TB is compared between the PSG ECG and ccECG signals, only a slight drop in AUC of about 1 % is observed. Substituting the TB by the Ramp EDR signal, causes only a

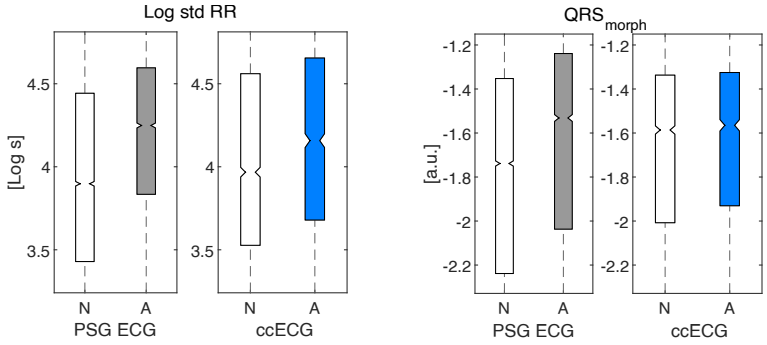


Figure 10.5: ECG features extracted from the Imec set for normal (N) and apneic (A) segments using the PSG ECG and ccECG.

slight drop in AUC of 3.4 % for the PSG ECG signal, whereas, for the ccECG signal this causes a drop of 12.7 %. Using the SR EDR estimation, a slightly worse performance is observed with the same AUC, but a sensitivity of only 49.9 % for the ccECG data. On the other hand, when comparing the PSG-based combination of PSG ECG and TB against the completely non-contact combination of ccECG and ccBioz, only slight changes in performance are observed.

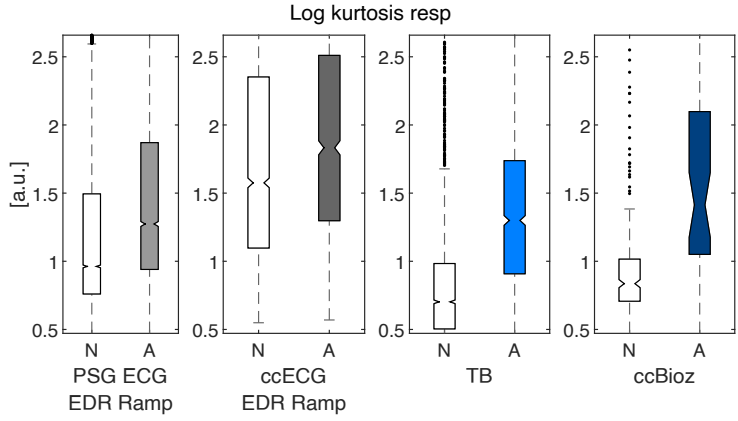


Figure 10.6: Log kurtosis of the respiration extracted from the imec set for normal (N) and apneic (A) segments using the different respiratory measurements.

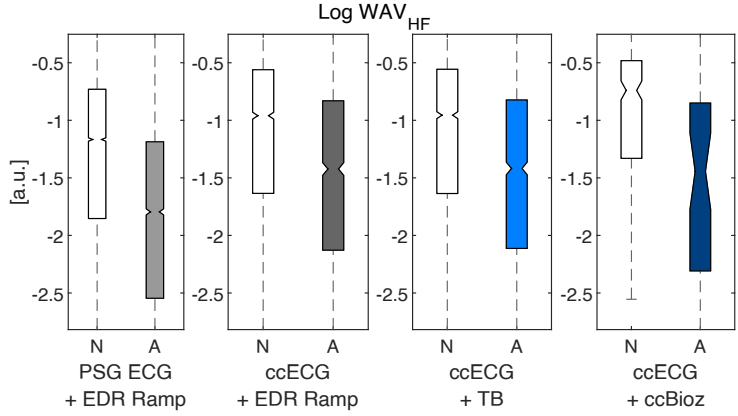


Figure 10.7: Log $\mathrm{WAV_{HF}}$ feature extracted from the Imec set for normal (N) and apneic (A) segments using different signal modality combinations.

Table 10.2: Apneic segment classification performance on the Imec data set for segments with clean capacitive sensor data.

Respiration	TB		EDR Ramp		TB	ccBioz	
ECG	PSG*	ccECG	PSG*	ccECG	PSG**	PSG**	ccECG
AUC [%]	81.2	**80.1**	77.8	*67.4*	81.1	80.4	**79.1**
Acc [%]	77.9	**77.2**	73.7	*63.5*	83.0	80.7	**80.4**
Se [%]	59.0	**60.7**	54.4	54.4	58.3	58.3	**66.7**
Sp [%]	86.2	**84.3**	82.0	*67.4*	87.5	84.8	**82.9**

* Including only segments with clean ccECG signal.

** Including only segments with clean ccECG and ccBioz signal.

Comparison of features extracted from PSG and capacitively-coupled measurements

Only three of the seven features of the original model show a weak to moderate correlation between features computed based on the PSG ECG + EDR Ramp and ccECG + EDR Ramp: the log std RR ($R^2 = 0.31$), WAV_{LF} ($R^2 = 0.12$) and log WAV_{HF} ($R^2 = 0.17$). The other features all have R-squared values below 0.05, meaning these features do not show the same characteristic behavior when extracted from ccECG in comparison to the contact ECG case. As described in [25], the most probable reason for this is that these features rely either on the complete ECG morphology or on the mechanical effect of respiration on the ECG for EDR extraction. As a consequence, when only features with an R-squared value larger than 0.1 were considered, all respiratory features were removed, pointing towards an inaccurate EDR extraction from the ccECG. Feature selection on the remaining features resulted in a set of five features, including the three features mentioned above, the log mean of the RR-intervals and the power in the HF band of the RR-intervals.

When a classifier was trained on the OSA+ data set using these five features, and tested on the ccECG + EDR Ramp signals, a slight increase in performance was observed. An AUC, accuracy, sensitivity and specificity of, respectively, 68.0, 67.5, 48.7 and 75.6 % were obtained. The sensitivity was, however, further decreased. This indicates the need of a good respiratory measure in order to obtain an acceptable sensitivity of the classifier. Better EDR estimates of the ccECG signal are thus needed, or if available, the ccBioz signal could be used as a measure of respiration.

Retraining the classifier based on capacitively-coupled data

The second classifier was retrained based on the ccECG data of 20 patients, selected based on their AHI, and tested on the other 8. This resulted in a test AUC, Acc, Se and Sp of, respectively, 63.9, 60.8, 54.9 and 63.0 % when using the Ramp EDR, which is worse than the results obtained in Table 10.2. Nevertheless, the training set performance achieved an AUC of 78.4 %, showing the potential of retraining based on the ccECG. When more ccECG training data would have been available, a better generalization might have been reached.

10.1.3 Discussion

Features

As mentioned in Chapter 7, the wavelet CR interaction features might be selected due to the fact that the wavelet decomposition acts as a filter which smooths out estimation errors in the EDR, while the other cardiorespiratory features take the unfiltered EDR signals as input. This would also explain the relatively high correlation between ECG and ccECG estimates for these features, as seen in Section 10.1.2, compared to the other respiratory features which all obtained R-squared values below 0.1. From these R-squared values, it is clear that the EDR-related features and the QRS_{morph} feature were not accurately estimated from the ccECG signal. This could be expected because these are largely dependent on the QRS morphology, while the ccECG signal can change between different leads from segment to segment and is obtained from different electrode locations than the PSG ECG. This causes the QRS morphology not to be consistent with that of the PSG ECG. In order to improve the EDR estimate from this ccECG signal, further research is needed related to the effect respiration has on the ccECG signals. The obtained results suggest that the ccBioz signals are a better source to extract respiratory features than the EDR from the ccECG signals. The possibility to have a fully unobtrusive apnea detection solution using ccECG and ccBioz measurements is therefore promising and should be further validated in a larger data set.

Apnea detection

In [135], the Ramp method was used to compute the EDR. This method relies on the amplitude of the R-peaks. In the case of ccECG signals, large amplitude shifts can occur. Therefore, the SR EDR estimate was proposed as an alternative. From the comparison in [132], it was concluded that this was one of the most

robust EDR methods. Our results, however, show a slightly better performance for the Ramp method on the ccECG. The fact that the SR method is still dependent on the QRS morphology might explain these results, further research into the effect that respiration has on the ccECG is needed. On the other hand, the ccBioz signal could be a good alternative to extracting the EDR from the ccECG which should be further investigated and tested on a larger population.

The results in Table 10.2 show a decreased sensitivity (for both PSG-based and capacitively-coupled based classification) when only segments containing clean capacitive data were considered. This can be explained by the fact that apneic events which have a clear cardiorespiratory reaction, due to an autonomic arousal, will be more likely to cause (movement) artefacts distorting the capacitive signals. These segments will thus be removed from the analysis due to the lower signal quality, and the apneic events left in the data set will have a more subtle cardiorespiratory reaction which is more difficult to detect. This explanation is supported by the observed apnea sub-type division, when considering the complete Imec data set, 64.9 % of apneic segments were linked to the less severe hypopnea events which only have a partial cessation of breathing. While for the subset with clean ccECG signals, this percentage increased to 74.9 %. In this study, segments with a low signal quality were discarded, but the periodicity and length of artefacts could also be studied in further research to investigate the link between apneic arousals and certain artefacts, as proposed in [62].

From Table 10.2 it was also concluded, that when the TB is used, the performance does not drop when the PSG ECG is substituted by the ccECG. However, further analysis has shown that when using the TB, the respiratory features were dominant in the classifier. When a classifier was retrained using only the three features extracted from the TB (std, kurtosis and PSD BW resp), still a performance with an AUC, Acc, Se and Sp of, respectively, 80.3, 77.9 57.6 and 86.7 % was obtained.

The fully unobtrusive apnea classification based on the ccBioz signal combined with the ccECG has shown good results, comparable to the ones obtained using PSG ECG and TB. These results are, however, based on a very small subset of the data set since quite some ccBioz data was lost due to sensor failure. This subset only contained 263 normal sleep minutes and 48 apneic minutes. Moreover, the ccBioz processing contained a manual artefact detection step. In order to obtain a fully automated method based on the ccBioz signal, an algorithm estimating the signal quality index should be developed. The obtained results using the ccBioz are promising, but should be further validated on a larger data set. New data set collections with the capacitively-coupled sensors are planned.

10.2 Apnea detection using a reflective Byteflies PPG sensor

In Chapter 8, an automated apnea detection algorithm based on the PPG signal was proposed. This algorithm was developed based on recordings from the gold standard in-hospital PSG during which the PPG signal was acquired using a transmissive finger clip pulse oximeter. This type of pulse oximeter is not ideal for home monitoring and can be uncomfortable to wear over multiple nights. Therefore, in this section, the use of a reflective Byteflies PPG patch attached to the wrist is investigated. Wearable Byteflies PPG signals were acquired simultaneously to the in-hospital PSG. The PPG-based OSA detection model from Chapter 8 will be tested on these recordings. As such, the usability and limitations of this wearable PPG sensor for the automated detection of OSA can be investigated.

10.2.1 Methods

Materials

Byteflies recordings were available for 100 subjects of the UZ Leuven OSA+ data set, and the 28 subjects of the UZ Leuven Imec data set. These recordings include three channels of PPG using, red, IR and green light recorded with a sampling frequency of 100 Hz. Additional information on the Byteflies sensor can be found in Chapter 2.

Signal synchronization

Before analyzing the wearable PPG signals, they had to be synchronized with the PSG data. Moreover, the sampling frequency of the Byteflies sensors was not exactly 100 Hz, and it could even differ from sensor to sensor. The data extracted from the sensor included only a starting time, no time stamps were available for the rest of the signal. Therefore, the exact sampling frequency should be defined. Otherwise a small inaccuracy in sampling frequency can lead to large delays when working with overnight recordings, which often last longer than 10 hours. In order to define the Byteflies sampling frequencies and to align the wearable PPG to the PSG, the PRV extracted from the PSG PPG and the wearable PPG were compared. An overview of the synchronization process can be seen in Figure 10.8.

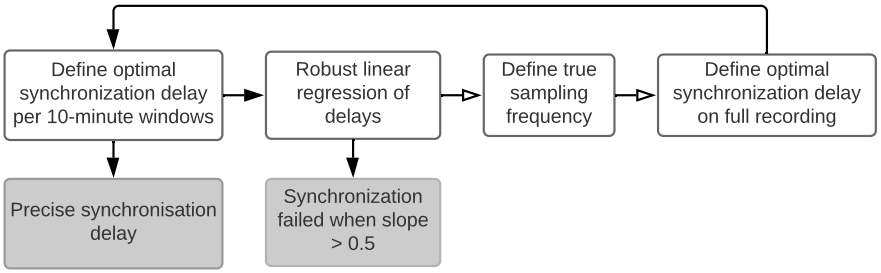


Figure 10.8: Overview of the steps in the synchronization of the wearable PPG signals to the PSG. The gray steps are only applied in the second iteration, which does not perform the white arrows.

First, the optimal synchronization delay is defined per 10-minute window. This is done by computing the cross-correlation of both PRV signals with different delays, the delay which yields the highest cross-correlation is then chosen as the optimal delay. Next, a robust linear regression was performed to model the obtained 10-minute window delays in function of the time for each patient and each PPG light frequency. Based on the slope of this regression, the true Byteflies sampling frequency could be obtained. In order to avoid that this sampling frequency estimation was distorted by artefacts, the median sampling frequency was computed over all light frequencies and all recordings acquired by the same sensor. It was thus assumed that the sampling frequency remained constant between recordings and that all light frequencies operated on the same sampling frequency. A sampling frequency of 99.296 Hz was obtained for the sensor used in room 505, and the sensor used in room 506 had a sampling frequency of 99.2914 Hz.

Now that the exact sampling frequency is known, the optimal synchronization delay was computed over the full recording using the cross-correlation as explained before. Additionally, a second iteration of the 10-minute window synchronization was performed, with a maximum shift of 1 minute. By taking the median delay over all segments, a more precise synchronization could be obtained. Finally, a robust linear regression was performed on the delays obtained using this 10-minute window synchronization. When the synchronization was successful, this regression should achieve a slope close to 0, meaning that the delay between the wearable and PSG PPG signals remains constant over all 10-minute windows. When the absolute value of this slope was larger than 0.5, the synchronization of the signals was assumed to have failed. This threshold was obtained by visual inspection of the synchronized signals.

In the end, out of the 100 OSA+ recordings, 87 could be synchronized. The first four recordings showed an unstable sampling frequency, which was solved after a software update. Whereas, for the other nine failed recordings the wearable PPG data was too noisy to extract a reliable PRV. For the Imec data set, one patient had no recording, and three others had too noisy wearable PPG signals. Only 24 recordings out of 28 could thus be synchronized.

PPG processing

The PPG processing and feature extraction was performed, as explained in Chapter 8. However, an additional Butterworth bandpass filter with cut-off frequencies of 0.3 and 5 Hz was applied. As such the baseline wander, and high frequency noise could be removed. Additionally, an amplitude threshold was put on the prefiltered signal. Segments where the absolute value of the PPG amplitude was larger than 4000 were removed. This threshold was defined empirically based on a visual check of the signals.

10.2.2 Results

PPG waveforms

In Figure 10.9 a 10 minute example of the synchronized and prefiltered PPG signals, and PRV estimates is given. Figure 10.10 zooms in on a 15 second window for the PPG signals, in order to get a better view on the PPG waveform. From these figures, it is clear that the Byteflies signals have different PPG waveform characteristics than the PSG PPG. This can be explained by the differences in measurement location (wrist vs finger) and oximeter acquisition mode (reflective vs transmissive). The red light channel of the Byteflies sensor had generally the most distorted PPG waveform, resulting in a worse PR estimation.

In the plotted PSG PPG signal from Figure 10.9 decreases in the PPG amplitude are visible, which might be linked to apneic events. These decreases, however, are not present in the Byteflies PPG channels. The extracted PR from these wearable measurements is similar as the gold-standard PSG PR. Especially the IR channel gave a good estimation for this example.

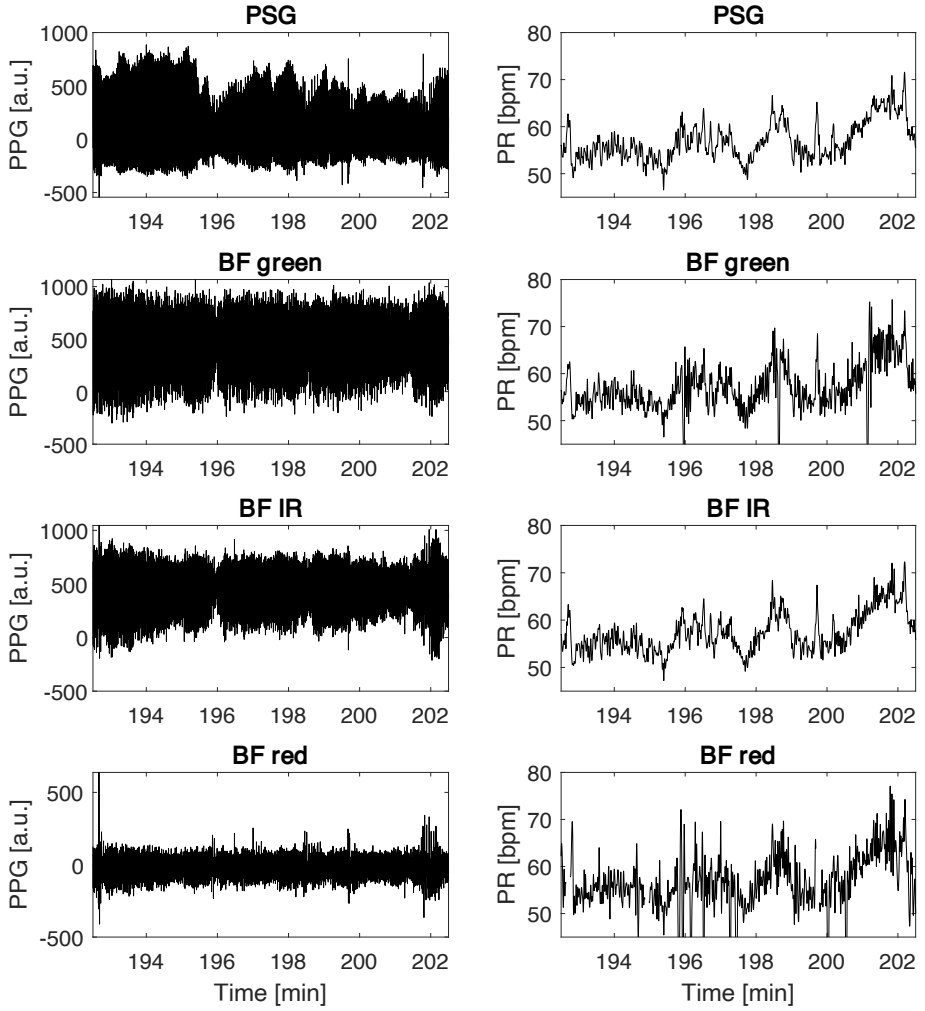


Figure 10.9: 10 minute PPG signal sample from the PSG and Byteflies recordings, with the extracted PR on the left.

Obtained PPG features

The features extracted from the IR channel of wearable PPG signals on the Imec data set for normal and apneic segments are plotted in Figure 10.11. When comparing these results with the obtained features for the PSG PPG in Figure 8.2, less clear differences in minimum PAV and std PRV are observed in the

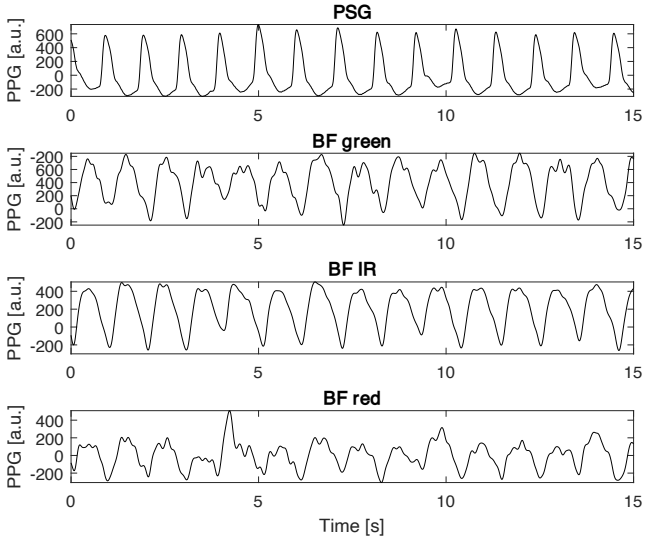


Figure 10.10: 15 second PPG signal sample from the PSG and byteflies recordings.

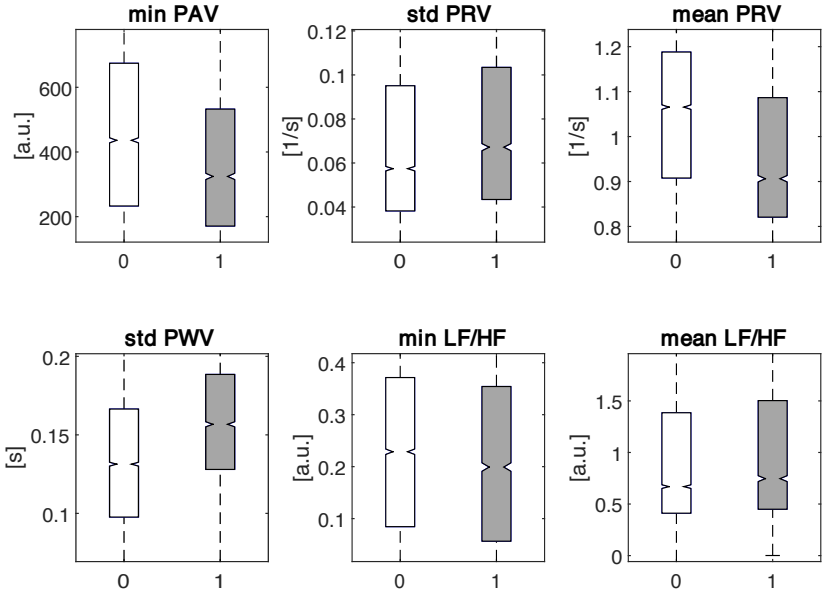


Figure 10.11: Extracted features from the IR channel wearable PPG signals of the Imec data set for normal breathing (0) and apneic segments (1).

wearable data set. The mean PRV, on the other hand shows a larger difference for the wearable data. Moreover, the range of values greatly differs between the PSG PPG and wearable PPG data for the min PAV and std PWV features. This can be due to the differences in signal acquisition and due to the fact that some artefacts might remain in the data set. Moreover, the wearable signal was prefiltered with a lower threshold on the lowpass filter, which might have altered the PPG waveforms. A threshold of 5 Hz was used instead of 35 Hz. This was needed to remove noise from the wearable signals.

Classification results

When the classification model, pretrained on the PSG PPG data set, was applied to the wearable data, sensitivities lower than 10 % were obtained. This is due to the difference in feature value ranges between the PSG and wearable data. Therefore, the classifier was retrained on the wearable data of the OSA+ set and tested on the Imec data set. The performance results of this retrained classifier on the Byteflies PPG of the Imec data set for each PPG light frequency are given in Table 10.3. Almost random classification occurred when using the red and green light frequency PPG channels, as AUCs around 50 % were obtained. Whereas, for the IR PPG channel, an AUC of 61.5 % was obtained. This performance is lower than the AUC of 69.2 % which was obtained on the PSG PPG for the Imec data set, however, these results are promising.

These preliminary results could be further improved by implementing better artefact detection algorithms and by a fusion of the different light frequencies. Moreover, the extracted features were selected based on the PSG PPG signals. However, the wearable PPG signal characteristics differ from the PSG PPG as they were acquired in the reflective mode compared to the transmissive one from the wrist instead of the finger. Therefore, a new feature optimization might be needed.

Table 10.3: Classification performance of the PPG apnea detection algorithm on the Byteflies PPG data of the Imec data set.

	Acc	Se	Sp	AUC	PPV
Red	50.9	53.1	49.9	51.6	32.2
Green	52.1	47.3	54.2	50.2	31.7
IR	60.1	57.3	61.4	61.5	40.0

10.3 Conclusion

In this chapter, the ECG and PPG algorithms developed based on PSG signals were tested on wearable measurements. The ECG algorithm was tested on ccECG and ccBioz data. When tested only on the unobtrusive ccECG data, a clear drop in performance was observed, which was linked to difficulties extracting the EDR from the ccECG due to morphological changes within the signal. The effect of the respiration on the ccECG should be further investigated in order to define better EDR measures. When the ccECG signal was instead combined with the TB or ccBioz, similar performances as for the PSG data set were obtained. Especially the results on the combination of ccECG with ccBioz are very promising, as this study was the first to test this fully unobtrusive solution. These results, should however, be further validated on a larger data collection of ccBioz signals.

The PPG algorithm was tested on wearable PPG data from the Byteflies sensor dots. The extracted features showed less clear differences between normal and apneic segments for the wearable data. Moreover, differences in the feature value ranges were observed. The apnea classifier thus needed to be retrained on the wearable data. For the IR light frequency, promising classification results were obtained on the Imec data set. Further research needs to be done on the artefact detection and signal fusion in order to improve these results.

Both experiments showed promising results, but also limitations of the current algorithms on the unobtrusive data. Therefore, the OSA-detection algorithms should be optimized for these unobtrusive signals. In order to do so, larger wearable data sets should be collected.

Part IV

Cardiovascular Comorbidities in Sleep Apnea

Chapter 11

Pulse oximetry markers for cardiovascular disease in sleep apnea

In this chapter, phenotyping of OSA patients beyond the AHI was investigated. The focus lies on the assessment of cardiovascular comorbidities in sleep apnea. The use of several pulse oximetry markers for the detection of CV comorbidities was investigated. The obtained markers are a major contribution of this PhD, they might give more physiological insights in why certain patients develop CV comorbidities, and might be used in the future to prioritize treatment for OSA patients at highest risk to develop CV problems. As such, the development of severe CV comorbidities might be averted.

This chapter is based on the paper presented at Computing in Cardiology 2019: ***Deviaene, M.,*** *Borzée, P., Buyse, B., Testelmans, D., Van Huffel, S., and Varon, C. Pulse oximetry markers for cardiovascular disease in sleep apnea. In 2019 Computing in Cardiology (CinC) (2019), IEEE, pp. P1-4.*

11.1 Introduction to OSA phenotyping

As discussed in the introductory chapter, using the AHI as measure for OSA severity has several drawbacks. There is a need for an improved phenotyping of sleep apnea in order to achieve a better patient selection and treatment

prioritization. In this study, the presence of cardiovascular comorbidities was chosen as patient differentiator. As such, the effect of sleep apnea on the cardiovascular system can be studied further. Moreover, the goal is to find markers in the non-obtrusive pulse oximetry signal in order to enable a wearable assessment, which could be brought to the home environment. Moreover, several studies have shown that oxygen desaturation severity parameters could be more predictive than the AHI as cardiovascular risk markers in OSA patients [7, 56, 70, 138]. The study by Sommermeyer et al. [120], used in addition to SpO_2, PPG parameters, which also had a statistically significant association with an increased cardiovascular risk. Combining pulse oximetry parameters from both the SpO_2 and PPG signals enables the assessment of the effect of both oxygen desaturations and arousals.

11.2 Materials

For this study, the UZ Leuven Matched, OSA+ and cluster data sets were used. These clinical data sets include a cardiovascular assessment of the patients at the time of the PSG. The patients were divided in three categories, Ct, cC and cE, based on their cardiovascular status as described in Chapter 2. The Ct group without any cardiovascular comorbidities and the cE group which experienced any cardiovascular event were used to select the most relevant cardiovascular markers. A classifier was trained for this 2-class problem, using a subset of the Cluster data set containing 45 Ct and 45 cE patients. The behavior of the intermediate cC group of patients suffering from any cardiovascular condition will also be discussed. In Table 11.1, an overview of patient demographics per data set and CV comorbidity class is given.

Only the SpO_2 and PPG signals of the PSG were used. These were recorded at 500 Hz for all data sets, but the Cluster set contains older recordings which were downsampled for archiving. Therefore, these signals were available at 200 or 50 Hz. For further processing, all SpO_2 signals were downsampled to 1 Hz and the PPG signals were resampled to 100 Hz.

11.3 Methods

11.3.1 Extraction of pulse oximetry markers

All oxygen desaturations were detected in the SpO_2 signals and 143 features were extracted as described in Chapter 6. These features include simple time-domain

Table 11.1: Patient demographics per data set and for each CV comorbidity class.

		# patients	Age (years) mean ± std	% Male	AHI (events/h) mean ± std
Training	Ct	45	47.7 ± 11.8	77.8	35.5 ± 12.5
	cE	45	65.3 ± 10.0	77.8	46.2 ± 20.3
Cluster	Ct	27	50.9 ± 7.9	77.8	31.2 ± 10.5
	cC	547	51.3 ± 11.4	80.9	40.8 ± 20.7
	cE	14	59.9 ± 7.4	85.7	33.9 ± 16.7
OSA+	Ct	46	41.5 ± 11.8	50.0	22.1 ± 26.1
	cC	140	51.5 ± 11.1	69.3	29.7 ± 23.7
	cE	11	61.7 ± 11.8	72.7	31.6 ± 21.0
Matched	Ct	50	47.3 ± 10.5	78.0	39.8 ± 23.3
	cC	44	47.4 ± 11.4	75.0	43.5 ± 20.6
	cE	6	54.2 ± 10.4	100.0	37.0 ± 25.5

(e.g. amplitude and length of desaturation), statistical (e.g. minimum and mean SpO_2 value) and desaturation severity (e.g. area below SpO_2 baseline) features computed on each desaturation. Moreover, novel quasi periodicity features (e.g. PRSA upslopes [10]) were computed on a 5-minute window around the desaturation. Afterwards, the feature values were averaged per patient for all desaturations with a minimal drop in SpO_2 of 2 %.

In a 120 second time window around each oxygen desaturation, the PPG peaks were detected in the signal and 9 PPG time-series were extracted as explained in Chapter 8. The PRV, PWV, PAV, STT, maximal pulse upslope, and area under the peak were extracted, as well as the instantaneous estimate of the power in the LF and HF bands and their ratio. The median value, standard deviation and difference between minimum and maximum value (delta) were computed for each of these time series. Additionally, the degree of cardio-respiratory interactions was estimated by the power in the respiratory component of the PRV, obtained using orthogonal subspace projections [131], taking the high pass filtered PAV and PWV as surrogates of the respiratory signal. The obtained PPG parameters were then averaged per patient, as done for the SpO_2 features.

11.3.2 Feature selection and classification

A total of 174 features were extracted from the pulse oximetry signals. In order to obtain an interpretable set containing the most meaningful features, the random forest-based feature selector proposed in this thesis and explained in Chapter 5 was applied.

Using the selected features, an SVM classifier with RBF kernel was trained in order to separate the Ct patients from the cE patients. Since age is a known risk factor for cardiovascular disease, and the cardiovascular groups have an increased age compared to cardiac controls (as can be seen in Table 11.1), the classifier was trained with and without including the age and/or AHI as an extra parameter.

11.4 Results

The proposed feature selection led to the selection of 3 SpO_2-based features, including: the minimal SpO_2 value, PRSA upslopes and ratio of amplitude of desaturation and resaturation. Additionally, 2 PPG features were selected, the standard deviation of the pulse upslope and the difference in STT. Boxplots of these feature values per data set, for the different cardiovascular categories (including cC), can be found in Figure 11.1. The Spearman correlation coefficient between these features and the age is, respectively, -0.25, -0.003, -0,17, -0,51 and 0,16.

Based on these features, different SVM classifiers were trained to assess the cardiovascular status of the patients (Ct vs cE), using features of one or both modalities, and with or without including the age. The performance results on the different test sets can be found in Table 11.2, this table also includes the results of SVMs based on the AHI and/or age.

11.5 Discussion

The resulting feature set shows more severe oxygen desaturations in patients with cardiovascular events (lower min SpO_2), a steeper upslope during the resaturation to baseline, and a larger amplitude of desaturation over the resaturation, which points to an incomplete recovery to baseline which might be followed by another SpO_2 drop. The PPG features point to a decreased variability of the pulse upslope, this parameter is highly correlated with the delta PAV ($C_{Spearman}$=0.84).

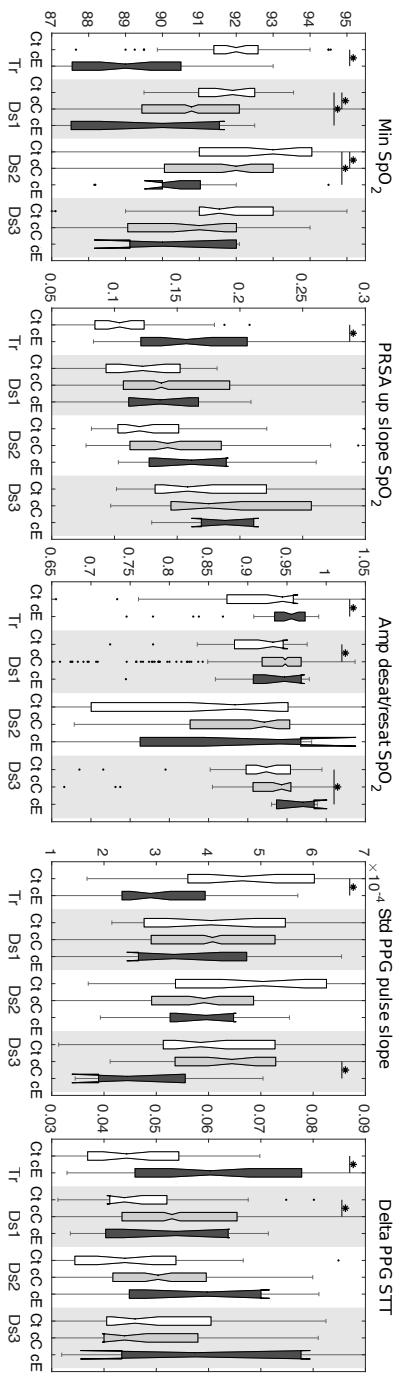


Figure 11.1: Overview of the selected features per data set (Training (Tr), Cluster (Ds1), OSA+ (Ds2) and Matched (Ds3)) and for each cardiovascular group. Significant differences between groups ($p < 0.05$) were marked with *.

Table 11.2: Cardiovascular comorbidity detection results between the Ct and cE classes.

	Training set			Cluster			OSA+			Matched		
	Se	Sp	AUC	Se	Sp	AUC	Se	Sp	AUC	Se	Sp	AUC
SpO_2	84.4	82.2	83.3	64.7	80.2	72.5	63.6	80.4	72.0	66.7	80.0	73.3
PPG	86.7	80.0	83.3	76.5	65.6	71.0	81.8	73.9	77.9	66.7	58.0	62.3
SpO_2 + PPG	82.2	82.2	82.2	64.7	78.1	71.4	63.6	78.3	70.9	66.7	78.0	72.3
AHI	53.3	64.4	58.9	23.5	71.9	47.7	27.3	82.6	54.9	16.7	62.0	39.3
Age	88.9	71.1	80.0	64.7	76.0	70.4	63.6	82.6	73.1	66.7	70.0	68.3
SpO_2 + Age	86.7	84.4	**85.6**	70.6	85.4	**78.0**	72.7	89.1	**80.9**	66.7	82.0	**74.3**
PPG + Age	86.7	75.6	81.1	64.7	76.0	70.4	63.6	82.6	73.1	66.7	70.0	68.3
SpO_2 + PPG + Age	84.4	84.4	84.4	70.6	83.3	77.0	72.7	84.8	78.8	66.7	82.0	**74.3**
AHI + Age	82.2	80.0	81.1	52.9	85.4	69.2	63.6	91.3	77.5	33.3	80.0	56.7

Since decreases in PAV have been linked to autonomic arousals [53], this could point to the fact that these patients have less arousals which could explain the more severe oxygen desaturations. The cardiac patients also have an increased difference in STT, this might be linked to the occurrence of irregular heartbeats. As can be seen in Figure 11.1, the features are quite consistent over the different data sets and the cC class lies in between the Ct and cE class as expected.

When studying the classification results in Table 11.2, it can be concluded that the classifier based on SpO_2 features including the age yields the best performance, an averaged AUC of 77.7% is achieved. The PPG features do not give any added value to this classifier. The age is a known cardiovascular risk factor and the results show that age is a clear confounding parameter, a classification AUC of 70.6% was achieved when taking only the age. However, adding the SpO_2 features caused a large improvement in performance. These parameters only have a weak correlation with the age. The PPG features, on the other hand, do not give any added value to the age-based classifier. This might be explained by the stronger correlation between age and pulse upslope standard deviation.

When taking only the AHI as predictor, quasi random classification occurred with an AUC of 47.3%. Adding the age to this classifier improved the AUC towards 67.8%, but this was still much lower than the performance achieved with the SpO_2-based features. Only for the OSA+ data set, acceptable results were obtained using the AHI and age, this is due to the fact that the cardiac control patients in this data set tended to have low AHIs, often <15, see Table 11.1.

In Figure 11.2, the posterior class probabilities of the SpO_2 and age SVM are compared between the different data sets and cardiovascular groups. A clear differentiation between the Ct and cE classes can be observed in all the data

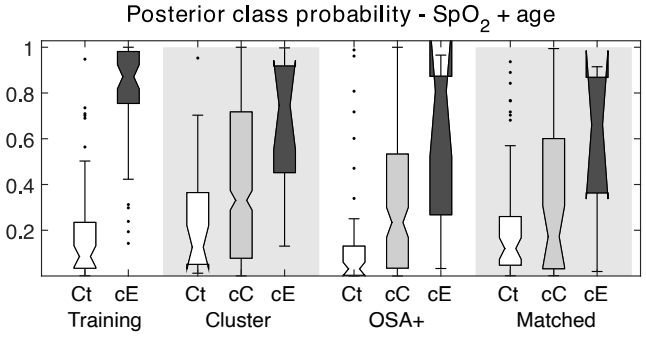


Figure 11.2: Posterior probability of the SVM classifier.

sets, the cC groups are situated in between those two classes, as expected.

11.6 Conclusion

The findings in this study suggest that SpO_2-based features can be of added value for the assessment of the cardiovascular status of sleep apnea patients. Patients with a cardiovascular comorbidity tend to have more severe oxygen desaturations and often do not have a complete resaturation to their baseline SpO_2 value. A decreased variability in the PPG pulse upslope, could point to less arousals in these patients. The investigated PPG-based features, however, did not show much added value to an age-based risk model.

Since the SpO_2 markers could be interesting for the phenotyping of OSA patients and their treatment decisions, they will be further studied in the next chapter. Moreover, a multilevel ICS model will be trained to assess the CV risk in OSA patients.

Chapter 12

Multilevel ICS to assess the cardiovascular status of sleep apnea patients using oxygen saturation markers

In this Chapter, the oxygen saturation markers defined in the previous Chapter for the assessment of the cardiovascular status of OSA patients are further investigated. In order to obtain an interpretable and automated clinical decision support system, a multi-level ICS model was designed to predict the CV risk based on these SpO_2 markers and patient demographics. Additional statistical tests were performed to analyze the differences between patient groups and the correlations between features. Moreover, the model was tested on two new data sets. The first one includes healthy volunteers, which was added to have a reference of the SpO_2 parameter values for subjects without any sleep disorder, nor any cardiovascular comorbidity. The second one includes CV follow-up data for up to 15 years, which makes it possible to test the predictive capability of the model

This study was published as: ***Deviaene, M.****, Borzée, P., Van Gilst, M., Van Dijk, J., Overeem, S., Buyse, B., Testelmans, D., Van Huffel, S., and Varon, C. Multilevel interval coded scoring to assess the cardiovascular status of sleep apnea patients using oxygen saturation markers. IEEE Transactions on Biomedical Engineering (2020).*

12.1 Materials

This study was performed using clinical PSG recordings of 974 subjects suspected to be suffering from sleep apnea, a control data set of 83 healthy volunteers, and a subset of 931 patients from the publicly available SHHS data set [108, 111]. The clinical data set (Leuven) consists of the UZ Leuven Cluster, OSA+ and Matched data sets, which were used in the previous Chapter. The healthy control data set (H) is the Kempenhaeghe data set, as explained in Chapter 2. In this study, only the 931 subjects of SHHS1 which had a follow-up PSG, but had no cardiovascular conditions at baseline were selected.

The SpO_2 signals were extracted from the overnight PSGs. These recordings had a length between 4.0 and 14.3 hours with an average recording length of 8.9 hours. The manually scored respiratory event annotations and hypnograms were available for each PSG. Moreover, the CV status of each patient was assessed at the time of the PSG, as discussed in Chapter 2. For the SHHS, the cardiovascular status at follow-up was taken as label, while the SpO_2 signals at baseline were investigated, in order to assess the predictive power of the developed model.

From the Leuven data set, 50 patients of each of the three CV categories were selected as training set (Leuven training) using k-medoids clustering on the age, BMI and gender [134]. The training subjects were selected as the centroids of the clusters. Using this approach, a training set is selected which represents better the underlying structure of the data set in terms of age, BMI and gender and any bias is avoided. An overview of the patient demographics for each data set and CV comorbidity category is given in Table 12.1.

12.2 Methods

An overview of the methodological steps applied in this study can be found in Figure 12.1. This section will first discuss the extraction of the oxygen saturation parameters. Next, the application of the multilevel ICS model to the assessment of cardiovascular comorbidities will be described together with the statistical methods used to interpret the results.

12.2.1 Oxygen saturation parameters

The sampling frequency of the available SpO_2 signals ranged from 1 Hz to 500 Hz, these were all downsampled to 1 Hz. Afterwards the signals were

Table 12.1: Patient demographics per data set for the cardiovascular control (Ct), condition (cC) and event (cE) groups.

			#	Age (years)	BMI (kg/m²)	AHI
				mean ± SD	mean ± SD	mean ± SD
Patients	Leuven training	Ct	50	44.7 ± 14.2	29.4 ± 6.0	31.1 ± 23.5
		cC	50	49.5 ± 13.7	34.8 ± 9.4	40.9 ± 25.9
		cE	50	63.0 ± 11.7	32.4 ± 5.6	42.5 ± 22.2
	Leuven	Ct	118	47.1 ± 9.7	28.2 ± 4.8	32.9 ± 20.3
		cC	679	51.2 ± 11.2	31.4 ± 5.9	38.6 ± 21.4
		cE	26	62.9 ± 7.1	30.3 ± 4.1	38.5 ± 18.0
General population	SHHS	Ct	346	54.4 ± 11.1	26.3 ± 4.2	11.4 ± 11.1
		cC	495	61.4 ± 8.5	27.7 ± 4.8	15.8 ± 13.4
		cE	90	68.5 ± 9.5	27.1 ± 4.3	15.5 ± 11.9
Healthy	H	Ct	83	33.7 ± 12.7	23.9 ± 3.1	1.3± 1.3

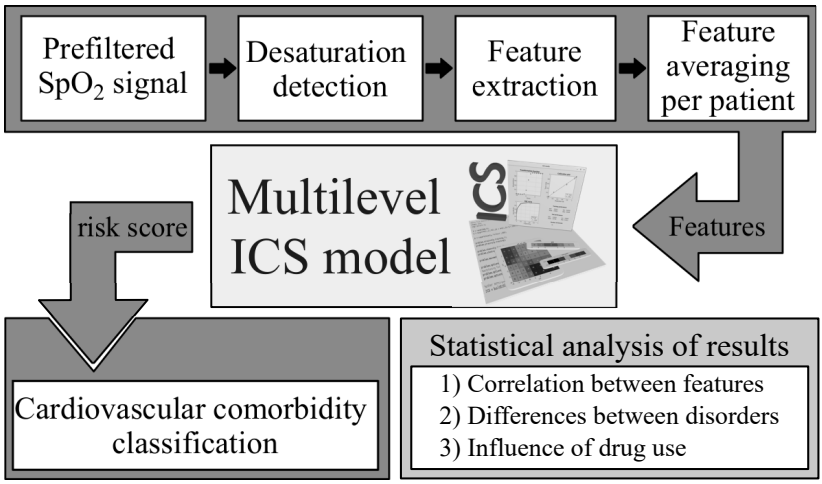


Figure 12.1: Overview of the applied methods.

pre-processed as described in Chapter 6. Next, the three parameters selected in Chapter 11, which formed the optimal parameter set to detect the presence of cardiovascular comorbidities, were extracted for each of the desaturations. These parameters are:

Minimum SpO_2 value The SpO_2 nadir has been shown to reach lower values for subjects at greater cardiovascular risk, as can also be seen from the example in Figure 12.2 on top.

PRSA upslope Using PRSA, quasi-periodicities were detected in the SpO_2 signal. The PRSA averaged resaturation was computed and the slope was extracted from this averaged curve [10]. The method was applied to 5-minute windows around each oxygen desaturation. First all upward points (anchor points, AP) were detected as can be seen in the middle plot of Figure 12.2, next segments of 10 seconds around each AP were extracted. These were then aligned, and the PRSA average was computed from these segments, as can be seen in the bottom plot of Figure 12.2. In [36], it was shown that OSA patients with a cardiovascular comorbidity have an increased PRSA upward slope, which can also be seen from the example in Figure 12.2.

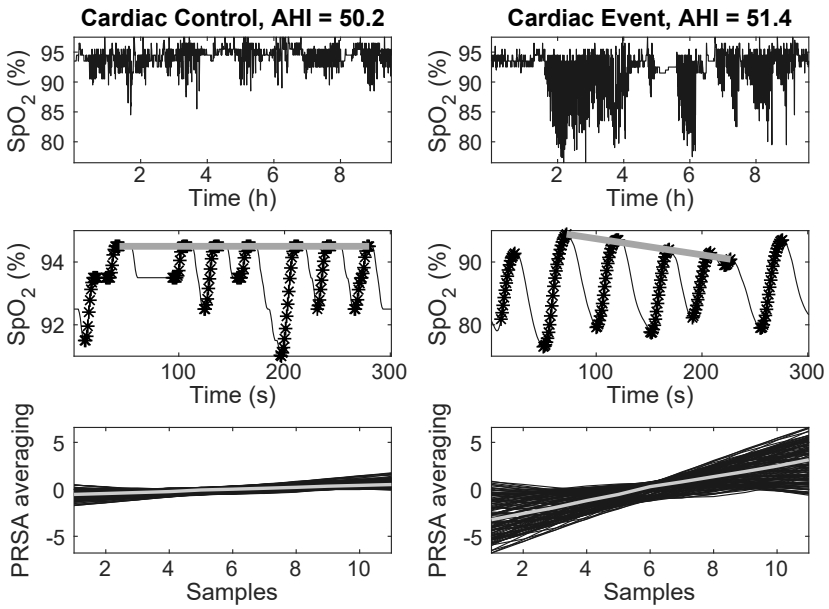


Figure 12.2: Comparison of SpO_2 parameters between a cardiovascular control and event patient with similar AHI. On top, the filtered SpO_2 signal is plotted, in the middle a 5-minute segment is shown with the PRSA anchor points indicated by '*'. Additionally, a thick gray line shows the SpO_2 baseline in the middle plots. At the bottom, the aligned PRSA slices are shown for this segment, together with the PRSA averaged curve in light gray.

Amplitude desat/resat The ratio between the amplitude of desaturation and the amplitude of resaturation is the last parameter. This parameter was computed by dividing the percentage of SpO_2 drop by the percentage that the SpO_2 increased again after the nadir. An increase in this ratio is seen for patients with a cardiovascular comorbidity, which points to incomplete resaturations to baseline. This can be seen in the middle plot of the example in Figure 12.2, where the SpO_2 baseline is plotted with a thick gray line. While the control patient has a stable SpO_2 baseline, the event patient has a decreasing SpO_2 baseline.

In order to obtain one value per patient, the median parameter value was taken over all desaturations with an SpO_2 drop from baseline larger than 2 %. In [36], the use of oxygen desaturation thresholds of 2, 3 and 4 % were investigated, and the most discriminative features were obtained with a threshold of 2 %. All desaturations in the recording were taken into account, independently of the sleep stage. The extracted parameters were then used to develop a classification model to assess the cardiovascular status per patient.

12.2.2 Multi-level ICS for the assessment of CV comorbidities

The multilevel ICS model, which was introduced in Chapter 4, was trained using the Leuven training subset to assess the CV risk in OSA subjects. The rest of the data was used as independent test sets. The training was performed with and without inclusion of parameter interaction terms and parameter preselection, the AUC cut-offs for model simplification were varied from 85 % to 97.5 %. In order to obtain easy-to-interpret models, the constraints of the mICS optimization were adapted to ensure the monotonicity of interval scores per parameter. mICS models were trained with different parameter sets including the derived SpO_2 parameters and patient information, such as age, BMI, gender and AHI. Moreover, an estimation of the AHI based on automatic processing of the SpO_2 signal by the algorithm provided in Chapter 6 [41] was also considered as input parameter for the mICS model.

The resulting mICS risk scores were then used to classify the subjects in the cardiovascular comorbidity categories. For this classification, the optimal mICS thresholds were defined as the point where the multiplication of sensitivity and specificity reached its maximum. This optimization was chosen in order to balance the sensitivity and specificity performance measures, so that subjects from the positive and negative class have a similar chance to be correctly classified.

Since the Framingham risk score [32] was available for 843 subjects of the Leuven data set, the results of the obtained mICS model were also compared

with a classification based on this risk score.

12.2.3 Statistical analysis

The obtained mICS risk scores and the extracted SpO_2 parameters were further analyzed using statistical tests on the Leuven data set. The correlations between the different parameters and patient demographics were quantified using the Spearman's rank correlation coefficient. This analysis was performed in order to study the influence that the patient demographics have on the extracted SpO_2 parameters.

Moreover, to have an idea about differences between groups of patients having different cardiovascular conditions the Kruskal-Wallis test with Bonferroni correction, with $\alpha = 0.05$, was performed between the different groups.

Many cC subjects were taking drugs for hypertension, diabetes or hyperlipidemia at the time of the PSG. In order to estimate the influence of these drugs on the extracted SpO_2 parameters, the differences between patients taking and not taking drugs were analyzed, per condition, using the Kruskal-Wallis test with $\alpha = 0.05$. For 100 patients of the Leuven data set, no information on drug usage was available, these patients were thus excluded from this analysis.

12.3 Results

12.3.1 Extracted SpO_2 parameters

Boxplots of the SpO_2 parameters per data set and per cardiovascular comorbidity group are shown in Figure 12.3. For the Leuven training and test set, the SpO_2 parameters follow the expected trends. Within the SHHS data set, however, the differences between the cardiovascular groups are not visible, and most values are within the range of the Leuven Ct groups. The healthy population (H) is clearly separated from the clinical data sets and for all parameters these differences are statistically significant.

The correlation between features was tested using the Spearman's rank correlation coefficient. Only for the min SpO_2 parameter a weak correlation with the age was found, with a correlation coefficient of -0.28. This correlation, however, dropped when it was computed within each cardiovascular category, a value of -0.20 was, for example, found for the cC group. The BMI, on the other hand, had a weak correlation with the AHI (0.29) and the three SpO_2 parameters; the strongest correlation was found with the min SpO_2 (-0.39).

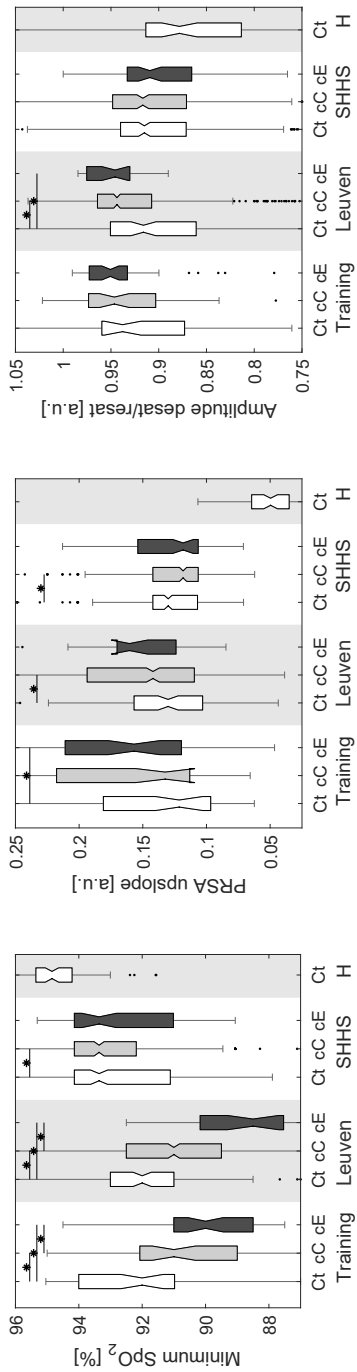


Figure 12.3: Overview of the extracted SpO_2 parameters for the different cardiovascular comorbidity groups (cardiovascular controls (Ct), conditions (cC) and events (cE)). Statistically significant differences between groups are marked with * $p < 0.05$.

The AHI also had moderate correlations with the SpO_2 parameters, having the strongest correlation with the PRSA upslope (0.60). Within the different SpO_2 parameters, the strongest correlation occurred between min SpO_2 and PRSA upslope (-0.64).

12.3.2 mICS Model

mICS models trained using different subsets of the extracted parameters and patient demographics were compared. When interaction terms were included in the model, they were removed during the model simplification step as they came out irrelevant. Therefore, further model optimization was performed without the inclusion of interaction terms. Since only 8 parameters were considered, the use of parameter pre-selection did not significantly reduce the computation time, nor increased the performance, therefore parameter pre-selection was not used. Different AUC cut-off values for model simplification were tested, setting this parameter to 97.5 % resulted in the best test set performances.

An overview of the training and test set performances to classify cC and cE subjects from Ct, on the different data sets for different subsets of features can be seen in Table 12.2.

When only age is considered, cC and cE subjects can be detected with a Leuven test AUC of 56.0 % and a sensitivity of only 17.9 %. If only cE and Ct subjects are considered, an AUC_{cE} of 74.0 % can be achieved. This model is very simple, since only two age intervals remained after the optimization procedure. As a result, patients older than 63 years were classified in the cardiovascular event group. However, due to the simplicity of the model, a poor generalization was achieved and no differentiation of mICS risk threshold could be observed between the cC and cE groups. The performance improved when the BMI was added to the model, an overall Leuven test AUC of 66.8 % was obtained and the sensitivity increased to 40.7 %.

The model trained only on the SpO_2 parameters has some predictive value with an overall test AUC of 61.1 % and an AUC_{cE} of 81.7 %. But adding the age to this model gave increases in performance of, respectively, 4 % and 9.5 %. Adding only the BMI to the SpO_2 parameters did not improve the performance. The best model performance was obtained when combining the SpO_2 parameters with the age and BMI. In this case, the overall Leuven test AUC increased to 69.5 % and cE subjects could be detected with an AUC_{cE} of 93.5 %. Next to AUC_{cE}, Table 12.2 also provides the AUC_{cC} for the classification between cC and Ct subjects, these are about 20 % lower than AUC_{cE} for the Leuven test set.

Table 12.2: Classification performance of the different mICS models and Framingham score for the detection of cardiovascular comorbidities (Ct vs cC + cE). AUC_{cC} and AUC_{cE} depict, respectively, the AUC of the separation between the Ct and cC classes, and the Ct and cE classes.

	Training					Independent Test Sets									
						Leuven					SHHS				
	Se	Sp	AUC	AUC_{cC}	AUC_{cE}	Se	Sp	AUC	AUC_{cC}	AUC_{cE}	Se	Sp	AUC	AUC_{cC}	AUC_{cE}
Age	33.0	90.0	61.5	52.0	71.0	17.9	94.1	56.0	55.3	74.0	45.0	76.3	60.6	58.4	73.2
Age+BMI	62.0	86.0	78.1	68.7	87.6	40.7	87.3	66.8	66.0	88.5	58.3	70.8	69.5	68.2	76.5
Age+BMI+AHI	44.0	88.0	73.7	65.6	81.8	25.5	92.4	64.1	63.4	82.7	39.7	84.4	65.3	64.3	70.7
SpO_2	48.0	82.0	70.2	63.2	77.2	37.9	83.9	61.9	61.1	81.7	28.4	82.9	52.9	53.7	48.6
SpO_2+Age	70.0	82.0	81.6	73.9	89.3	45.0	82.2	65.9	64.9	91.2	46.8	76.3	65.2	64.3	70.0
SpO_2+Age+BMI	**72.0**	**78.0**	**82.8**	**74.9**	**90.7**	**47.5**	**80.5**	**69.5**	**68.6**	**93.5**	**44.6**	**78.3**	**67.1**	**66.3**	**72.0**
SpO_2+Age+BMI+AHI	72.0	78.0	83.0	75.5	90.4	48.5	80.5	69.0	68.1	93.3	47.9	76.3	68.9	67.5	76.9
AHI	100.0	14.0	57.0	57.0	57.0	97.7	5.1	51.4	51.4	52.5	75.2	39.0	57.1	56.8	59.0
AHI SpO_2	96.0	22.0	59.0	59.0	59.0	93.5	11.0	52.2	52.1	55.5	81.2	32.4	56.8	56.4	59.0
Framingham	65.0	88.0	81.8	77.0	87.1	52.8	92.4	80.9	80.4	94.9	Not available				
Framingham+mICS	69.0	88.0	84.6	78.6	91.1	53.2	90.7	81.1	80.5	96.9					

Classifying subjects based on the AHI or the SpO_2 based estimate of the AHI gave poor results, AUCs of, respectively, 51.4 and 52.2 % were obtained. Adding the AHI to the models mentioned above also did not improve performances. In the last rows of Table 12.2, the classification results based on the Framingham risk score are presented for the Leuven data sets. An optimal risk score cut-off value of 12.2 was obtained based on the training data. With a test AUC of 80.9 %, classification based on the Framingham score outperforms our proposed model, especially when focusing on the cC subjects. But the computation of the Framingham score already includes some of the cardiovascular conditions; the presence of diabetes is included as well as the cholesterol and blood pressure levels. Moreover, if the necessary parameters would be available to compute the Framingham score, it can be combined with the proposed mICS risk model. Classification based on the multiplication of the mICS risk score and the Framingham score resulted in an increase in overall test AUC to 81.1 % and a AUC_{cE} of 96.9 %.

Table 12.2 also includes the performance scores on the SHHS data set, a clear drop in sensitivity and AUC can be seen for the SpO_2 based model. This could be expected, since the SpO_2 features did not show any differences in Figure 12.3.

The final model including SpO_2 parameters, age and BMI is shown in Figure 12.4. All three SpO_2 parameters came out the mICS model optimization as relevant and were included in the final model.

The resulting mICS risk scores for the different cardiovascular groups and data sets are shown in Figure 12.5. A clear separation can be seen between the three

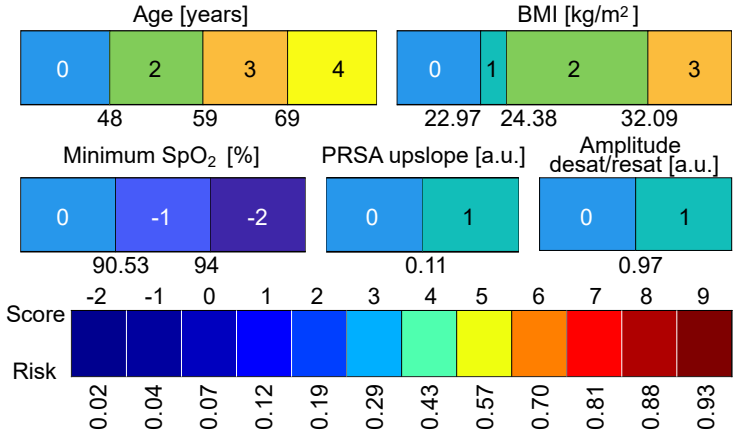


Figure 12.4: Resulting mICS model including SpO_2 parameters, age and BMI.

groups for the Leuven training and test sets. The Ct, cC and cE groups have, respectively, median values of 0.29, 0.43-0.57 and 0.70-0.81. For the SHHS data set the Ct and cC values lie within similar ranges as the Leuven data, but the cE group does not have an increased risk with respect to cC. This lack of increased risk could also be extracted from the similar AUC_{cC} and AUC_{cE} in Table 12.2. The H group has very small risk scores, with a median value of 0.07.

The correlation between the mICS risk score and the Framingham risk score was tested. A correlation of 0.66 was obtained, which was mainly due to the fact that both risk scores highly depend on the age.

The resulting confusion matrix for 3 class classification on the Leuven test set is plotted in Figure 12.6. Again a clear separation between the Ct and cE groups is seen, but the cC overlaps with both groups.

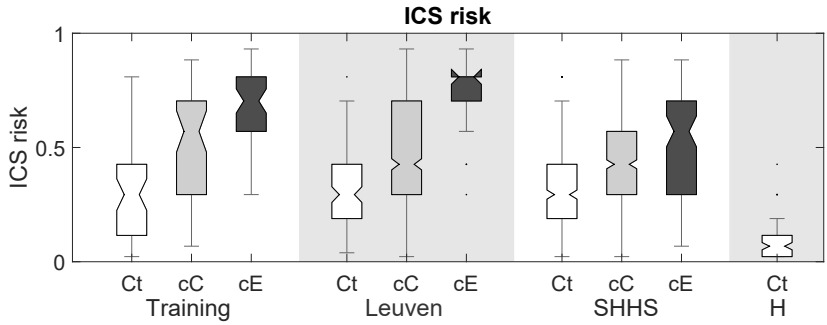


Figure 12.5: mICS model risks for the different data sets and cardiovascular groups; control (Ct), condition (cC) and event (cE).

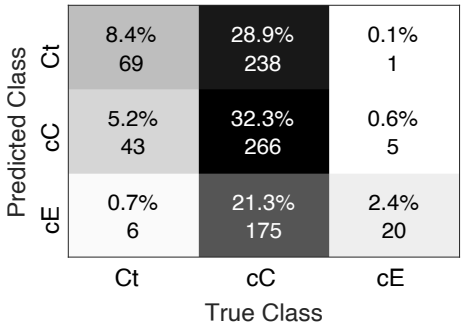


Figure 12.6: Confusion matrix of classifying Leuven test subjects using the mICS model including SpO_2 parameters, age and BMI.

12.3.3 Differences between cardiovascular disorders

The obtained mICS risk scores were compared between the different cardiovascular disorders in order to investigate whether differences between disorders could be seen. Firstly, the cardiovascular conditions were compared, 6 groups of patients were formed: 156 subjects which only had hypertension were compared with 120 with only hyperlipidemia, 307 which had both hypertension and hyperlipidemia, 95 subjects which had diabetes amongst other conditions, 17 subjects with arterial disease and 19 with atrial fibrillation (AF). Statistically significant differences in mICS risk scores were found between these groups of patients. The diabetic and AF patients, on the one hand, had an increased risk correlating to a statistically significant decrease in min SpO_2. The other two SpO_2 parameters also displayed a tendency towards higher values for these patients, but these differences were not statistically significant. Diabetic patients also had a significant increase in BMI with respect to the other groups, whereas the AF population had an increased age. The patients with hyperlipidemia, on the other hand, had a decreased mICS risk, which was caused by a statistically significant decrease in age. The SpO_2 parameters did not differ between the hyperlipidemia, hypertension and hyperlipidemia plus hypertension groups. Finally, a tendency towards an increased risk for the arterial disease group could be observed, which was caused by an increase in age.

Within the cardiovascular events, 4 groups could be compared: 36 subjects had a myocardial infarction, 19 a heart failure, 15 a stroke and 6 subjects had more than one type of event. No statistically significant differences could, however, be observed between these four groups.

12.3.4 Influence of drug use

The influence of drug use on the mICS risk score model was assessed by comparing those who did not take cardiovascular drugs to those with the same disorder, who were taking drugs. A statistically significant increase in age was found for both hypertension and hyperlipidemia patients taking drugs. This increase was also observed for patients suffering from both disorders which were taking drugs for either disorder. These age increases resulted in an increased mICS risk. For the hyperlipidemia group, no statistically significant differences in SpO_2 parameters could be observed between both groups. Subjects taking anti-hypertensive drugs, however, had a statistically significant decrease in minimum SpO_2 values, independently of the fact whether they were taking drugs for hyperlipidemia. This analysis could not be performed for anti-diabetic drugs, since in this group many patients were taking multiple types of drugs.

12.4 Discussion

This study developed an interpretable risk score model to assess the cardiovascular status of OSA patients based on SpO_2 parameters and patient demographics. The extensions to the ICS model, proposed in Chapter 4, in which multilevel classification is possible and monotonicity of the mICS interval scores per parameter is imposed, were applied. These extensions can be generally used in other applications including different levels of severity.

The novel SpO_2 parameters, proposed in the previous chapter for the assessment of cardiovascular comorbidities in OSA patients, and the obtained results, were further analyzed using statistical tests. A set of healthy recordings was added to the analysis in order to get a reference on how the values of these SpO_2 parameters behave in the general population. The model was also applied to a subset of the SHHS in order to have an idea on the predictive capabilities of the mICS model. This section will further discuss all of the obtained results.

12.4.1 SpO_2 parameters and cardiovascular comorbidities

The correlation analysis of the SpO_2 parameters and the patient demographics revealed that the minimum SpO_2 value was mainly correlated with the BMI. This relationship also became clear in the analysis of the different conditions, in which patients having diabetes had both an increased BMI and decreased minimum SpO_2 values. This relationship between the BMI and the severity of oxygen desaturation was previously shown by Peppard et al. [105]. The PRSA upslopes, on the other hand, were more correlated with the AHI. This relationship can be expected since the PRSA searches for quasi-periodicities in the SpO_2 signal, and the PRSA averaged curve will be more pronounced if multiple oxygen desaturations occur within the 5-minute segment under investigation.

The analysis on the influence of drug use showed a statistically significant decrease in minimum SpO_2 for patients taking anti-hypertensive drugs. This group of patients, however, also had a non-significant increase in BMI, which could have affected the minimum SpO_2 instead of the drug use, since the BMI is correlated with the minimum SpO_2. Therefore, no conclusive effects of drug use on the SpO_2 parameters, independently of changes in patient demographics were found in this study. Another conclusion that came out of this analysis, is that patients treated for cardiovascular conditions tended to be older. This is in accordance to the Hypertension and Diabetes Risk Screening and Awareness study [118], which observed that older hypertensive patients tended to be more often diagnosed and treated than younger hypertensive patients. The severity

of the disorder, measured by blood pressure and blood analysis, could not be linked to this difference, since these measurements were only available at the time of the PSG, when patients were already taking drugs. Therefore, patients on drugs had a lower observed cardiovascular disease severity.

Studies have shown that OSA patients have more severe oxygen desaturations during rapid eye movement (REM) sleep [75]. The presented results were, however, obtained by averaging the SpO_2 parameters over all desaturations, independent of the sleep stage. An additional analysis of the influence of sleep stages, nevertheless, had shown no increase in performance when REM and non-REM sleep were separated. It is thus not necessary to differentiate between REM and non-REM desaturations and no sleep staging is required for the algorithm. Additionally, the ratio of the number of REM over non-REM events was also not significantly different between the cardiovascular comorbidity categories.

12.4.2 mICS model for cardiovascular comorbidity assessment

The final mICS model obtained an overall Leuven test AUC of 69.5 % for the differentiation between patients without any cardiovascular disease Ct and those with a cardiovascular comorbidity. Adding the SpO_2 parameters gave an improvement of 2.7 % compared with the model using only age and BMI. For the detection of cE patients, an improvement in performance of 5 % was obtained, whereas for the detection of cC patients, the increase in performance was only 2.6 % in AUC_{cC}. This improvement in performance might seem limited, but the SpO_2 parameters showed a clear correlation with the cardiovascular comorbidity classes for the Leuven data set in Figure 12.3, while only weak correlations between these parameters and age or BMI were found. These results indicate that the SpO_2 parameters have an added value compared to the age and BMI. Since the cardiovascular comorbidity classes are not matched for age and BMI, the obtained model is highly dependent on those parameters. If these mICS models would be tested on a more balanced population in terms of age and BMI, the SpO_2 parameters might provide larger increases in performance. From a cardiovascular prevention view, the focus would ideally go to the youngest OSA subjects. In this sub-population it is, however, hard to find subjects with detectable cardiovascular comorbidities.

The age dependency of the model also leads to the fact that, in general, misclassified Leuven Ct patients were significantly older than correctly classified Leuven controls. Especially, it should be noted that subjects older than 69 years and with a BMI higher than 32.1 have a minimum mICS score of 5, resulting in a predicted risk of 57%. Therefore, all these subjects were classified in the

cE group. In order to have a better differentiation within this demographic group, more Ct patients should be included in this range. On the other hand, the model will always classify subjects younger than 48 years and with a BMI smaller than 24.4 in the Ct group. For this age and BMI range, no cE subjects were available. The performance to classify cE from Ct subjects was recomputed for all subjects lying outside these two demographic ranges, an overall Leuven test AUC of 67.2 % was obtained, with an AUC_{cE} and AUC_{cC} of, respectively, 92.3 and 66.2 %. Also an increase in AHI can be seen for misclassified controls and they tended to reach lower SpO_2 values.

Moreover, when analyzing the signals of 7 Ct patients in the Leuven data set, a large number of ectopic beats could be observed in the electrocardiogram signals of these patients. Although these patients were labeled as cardiovascular controls, they could already have been far in the development of a cardiovascular comorbidity, which was still unknown at the moment the PSG was recorded. These subjects had mICS risk scores ranging from 0.12 to 0.70, with a median value of 0.57, while the Ct group had for this data set only a median risk of 0.29. An increased cardiovascular risk could thus already be detected.

There were 6 cardiovascular event patients misclassified, one myocardial infarction patient was classified in the Ct group. The other five patients were classified in the cC group, two had a myocardial infarction, two heart failure and one subject had a stroke.

12.4.3 Predictive power of the obtained mICS model

A drawback of the Leuven data sets, is the fact that the patients only had an assessment of cardiovascular comorbidities at the time of the PSG. No follow-up information was available, this means that cardiovascular control patients could have developed a cardiovascular comorbidity shortly after their PSG, as could have been the case for the 7 controls with ectopic beats as discussed above. Therefore, the publicly available SHHS was included in this study. This test had an overall AUC of 67.1 % to predict whether cardiovascular control subjects were going to develop a cardiovascular comorbidity or not. This performance is, however, lower than the one on the Leuven set. Moreover, when the SpO_2 parameters were investigated, no clear separation for the SHHS groups could be observed. Especially, when looking to the minimum SpO_2 parameter, for which the three cardiovascular groups, Ct, cC and cE had median values of 93 %. This parameter gave the clearest separation for the Leuven data set, but for the SHHS it does not provide any separation. These values are all in the range of the Leuven Ct groups. This points to the fact that the subjects included in the SHHS have less severe apnea related oxygen desaturations, and thus have

no increased cardiovascular risk reflected on the SpO_2 parameters. This could be due to the fact that for this test set, the SpO_2 parameters were computed at the baseline PSG. At this moment, no CV comorbidities were present in any of the patients. Only during the follow-up period, of up to 15 years after this baseline PSG, cC and cE subjects developed their CV comorbidity. The SpO_2 parameters might therefore still be similar between the different groups. The first cardiovascular event could have occurred up to 15 years after the PSG which was used to extract the SpO_2 parameters. It was tested whether limiting the time between the PSG and the occurrence of cardiovascular events improved the results. No significant differences were, however, observed when only the 24 cE subjects were included which had a cardiovascular event within 5 years of their baseline PSG.

It should also be taken into account that the SHHS data set contains subjects from the general population, which often did not have any sleep related complaints. Even when the population with an AHI larger than 15 is considered, the minimum SpO_2 values do not decrease. Moreover, the mean ESS [66] is only 7.7 for this subgroup, while only a score above 10 is considered abnormal. For comparison, the average ESS for the Leuven data set is 10.4. Therefore, the developed cardiovascular comorbidities might be less related to the sleep apnea events in the SHHS than in the Leuven population, thus resulting in no abnormal SpO_2 parameters. Additionally, the SHHS subjects are significantly older than the Leuven population. The average age at baseline was 54.4 years for the controls of the SHHS, which is 10 years older than in the Leuven data set. This will further complicate the differentiation between the cardiovascular groups since studies suggest that although the overall cardiovascular risk increases with age, the cardiovascular risk associated with OSA decreases with age [55]. This decreased association might be explained by a healthy survivor effect, in which OSA subjects with a high cardiovascular risk are unlikely to be alive or without any cardiovascular condition from certain ages on.

These results show that one should be careful with diagnosing OSA based on AHI alone, in order to avoid over-diagnosis and over-treatment of patients without complaints that might not have an increased long-term risk. The proposed parameters could therefore assist to prioritize treatment for those patients which have more severe SpO_2 features, linked to a higher cardiovascular risk.

12.4.4 Influence of training set selection on the performance

The training set was selected based on the age, BMI and gender of the subjects. It was tested which performances could be obtained when the selection of training subjects was performed at random. 10 iterations were run, and the detection AUC on the Leuven test set ranged from 59.5 to 70.4 %, with an average of 68.8 %. Moreover, for the AUC_{cE} a maximum of 90.1 % was obtained, while our classifier obtains an AUC_{cE} of 93.5 %. This shows that the intelligent selection of training samples using k-medoids clustering increases the performance of the model. Including subjects from the SHHS in the training set resulted in a decrease in performance, so this was not considered further.

Moreover, 10-fold cross-validation was tested on the Leuven data set. Within each fold, an equal amount of Ct, cC and cE subjects were selected for training using the k-medoids algorithm, in order to ensure class balance. An averaged AUC and AUC_{cE} of, respectively, 67.0 and 85.8 % were obtained when using the age, BMI and all SpO_2 features. When only taking into account the age and BMI these performances dropped to 65.1 and 81.5 %. These results have a drop in performance, but show similar trends when compared to the results presented in Table 12.2 using a fixed test set. Additionally, when using cross-validation slight changes in the model were observed between different folds. Therefore, the manuscript was focused on the final ICS model obtained with a fixed training set. If more data would be available, the cross-validation models would become more stable.

12.4.5 Limitations of the study

One of the drawbacks of the used Leuven data set is the lack of patient follow-up, as discussed above. Nevertheless, this data set represents the patients that are referred to the sleep lab, which have OSA-like complaints. This is the most interesting population to develop a cardiovascular phenotyping of OSA patients. Since studies have shown that symptomatic OSA patients have a higher risk to develop cardiovascular events. Whereas, in general population data set, as the SHHS, subjects can have a-symptomatic OSA with an AHI larger than 15 and a cardiovascular problem, but these two problems might be unrelated [70, 48]. Moreover, future follow-up studies could also take into account cardiovascular conditions and treatments that are present at baseline and track if the cardiovascular comorbidities worsen. When doing so, the comorbidities and treatments present at baseline could be added as input to the classification model.

Despite the large number of recordings that were available for this study, there was a large gap in patient demographics, such as age and BMI, between cardiovascular control and event patients. Patient matching would have greatly impacted the size of our data set, and this data set represents the patient population that is recorded at the sleep lab in clinical practice. Therefore, it was decided not to correct for patient demographics, but to take them into account in the definition of the risk score. In order to obtain a model that is useful in daily clinical practice. The H subjects used as reference also differed substantially in age and BMI from the Leuven data sets.

As future work, the addition of novel PPG features could be investigated as well as new measurement and artefact detection techniques to improve the reliability of PPG features. The mICS model could be compared with other interpretable classification models as well. Moreover, if more data would be available, an indepth analysis towards the influences of cardiovascular diseases and drug use on the risk score could be performed.

The obtained cardiovascular risk model for OSA patients showed good performances, but should be tested further on clinical data sets with cardiovascular follow-up, before it can be used in clinical practice for OSA phenotyping. Since the used parameters are extracted from the SpO_2 signal, which can be easily acquired at home, multiple night testing could also be investigated in the future.

12.5 Conclusion

This study showed that parameters based on SpO_2 are of added value to predict the cardiovascular comorbidity status of OSA patients. A multilevel interval coded scoring model was constructed based on age, BMI and these SpO_2 parameters. This model was able to predict whether a subject had any cardiovascular comorbidity with a test AUC of 69.5 %, and if (s)he suffered from a cardiovascular event or not with an AUC of 93.5 %. The resulting mICS risk scores of cardiovascular condition patients showed an increased risk, but lower than those of subjects who experienced a cardiovascular event. This model could be useful in clinical practice to estimate the cardiovascular risk of sleep apnea patients, and to do a screening to prioritize patients for treatment.

The SpO_2 parameter extraction and mICS risk score computation presented in this chapter was included in the Python GUI described in Appendix A.

Chapter 13

Conclusions and future directions

In the introduction, two research goals for this thesis were defined: to develop a reliable, multimodal OSA screening method, and to define markers for the characterization of OSA patients beyond the AHI. This final chapter will discuss the main findings and contributions made in this thesis to both of these goals and to the algorithm development. Moreover, some open problems and suggestions for future work will be presented.

13.1 Conclusions

Obstructive sleep apnea is the most common sleep related breathing disorder. However, subjects suffering from it often remain undiagnosed. Moreover, the prevalence of OSA is increasing and a better phenotyping of patients is needed to select a personalized treatment.

The goal of this thesis was to tackle these two challenges in OSA diagnosis:

1. To develop a reliable, multimodal screening method based on unobtrusive measurements such as the SpO_2, ECG, PPG and respiratory measures.
2. To define markers for the characterization of OSA patients beyond the AHI, based on the presence of cardiovascular comorbidities.

In the next sections, the main contributions of this thesis on both of these goals will be discussed. Moreover, in Part II, algorithmic contributions were presented, and an overview of these contributions will also be included.

13.1.1 Multimodal OSA screening methods

The first objective of this thesis was to develop multimodal OSA screening methods based on unobtrusive measurements. Chapters 6 to 8 focused on the development of unimodal algorithms, which were then combined in Chapter 9. These methods were also tested on wearable measurements in Chapter 10.

The first unimodal algorithm presented in Chapter 6 was based on the SpO_2 signal. Features were extracted from oxygen desaturations, including novel quasi-periodicity estimates. These were then used to classify if the desaturations were caused by apneas using a random forest classifier. Robust linear regression was subsequently used to estimate the AHI from the number of detected apneic desaturations per hour of recording. The model was tested on different independent test sets including recordings from both the general population and subjects suspected of suffering from OSA. Good performances were obtained over the different test sets. However, the results improved when the model was retrained on the specific type of population. When applying this algorithm, it should thus be taken into account if the goal is to screen in the general population or in a clinical group of patients.

In Chapter 7, an ECG-based sleep apnea detection algorithm was developed which builds further on the work presented by Varon et al. [135]. The feature set was extended with additional cardiorespiratory measures and the slope range was considered as an alternative EDR signal. Moreover, the algorithm was designed such that the EDR signal could easily be replaced by other respiratory measures. Unfortunately, the extra cardiorespiratory features were not selected in the final feature set, and the slope range EDR did not yield a better performance than the Ramp EDR. The fact that both EDR signals obtain similar performances, might be linked to the nature of the data sets. During sleep the changes in dynamics, and the number of artefacts are limited. Thus the use of more complex EDR methods might not be needed. The final model obtained similar results as in [135] on the same data set. However, on the more recent UZ Leuven data sets, the AUC dropped with almost 10 %. This shows that the performance of this algorithm is highly dependent of the used data set, even when the model was retrained. The SpO_2-based model, on the other hand, was more stable over different data sets.

A PPG-based apnea detection algorithm was presented in Chapter 8. This algorithm made use of PPG time series which are modulated by the respiration

and/or the autonomous nervous system. Moreover, a point-process model was used to estimate the instantaneous powers in the HRV frequency bands. This model showed promising results, but performed the worst of the unimodal OSA detection algorithms. Especially, low positive predictive values were obtained using the PPG-based model. The occurrence of a high number of false positive detections in PPG-based OSA screening models is a common problem in the literature. These false positives are probably caused by arousals with other origins than apneic events.

Chapter 9 presented different multimodal experiments for OSA screening. First, the thoracic belt respiratory effort signal was included in the ECG-based algorithm instead of the EDR. This led to a clear improvement in performance. Next, the added value of the extracted ECG and PPG features to the SpO_2-based classifier was investigated. These modalities were merged in a straightforward way by feeding the combined feature sets, extracted from 1-minute windows, to an LS-SVM classifier. However, no clear improvement on the SpO_2-based classification performance could be obtained. In the next experiment, a "smarter" merging of the SpO_2 and PPG features was considered by designing a 4-state hidden Markov model. However, the addition of a wake and apnea response state did not improve the classification performance. Moreover, the state sequential information did not bring any added value to the model. The last experiment of Chapter 9 investigated the AHI estimation and OSA screening performance of the unimodal and multimodal algorithms. Moreover, the impact of removing wake periods on the results was evaluated. This analysis showed that only the SpO_2-based algorithm is able to obtain AHI estimations and OSA screening performances which are useful in clinical practice. The multimodal algorithms could not improve these performances. The removal of wake periods had only a small effect on the AHI estimation, which in the end did not alter the OSA screening performance.

Finally in Chapter 10, the ECG and PPG algorithms were tested on wearable measurements. The ECG algorithm was tested on ccECG and ccBioz data. The model only including the ccECG showed a decreased performance. This was linked to the fact that the extracted EDR did not reflect changes due to respiration in the ccECG signal, as the physical measurement principals differ between the ccECG and the ECG. When the ccBioz was added as respiratory measurement, similar performances as on the PSG data were obtained. However, these results should be further validated on larger data sets including ccBioz signals. The PPG algorithm was tested on wearable PPG data from the Byteflies sensor dots. On the IR light channel, promising classification results were obtained. But better artifact detection algorithms should be implemented and the feature set should be optimized for these wearable PPG signals, in order to obtain comparable results to the PSG-based experiments.

The novel SpO_2 model turned out to be the best performing OSA screening method. The multimodal models which also included ECG and PPG features could not improve this performance. Smarter ways are thus needed to merge these modalities. Promising results were obtained on the unobtrusive measurements, but the algorithms should be further tested on larger data sets and the PSG algorithms should be further optimized for these wearable signals.

13.1.2 Characterization of OSA patients beyond the AHI

The next part of the thesis focused on the characterization of OSA patients beyond the AHI. The presence of cardiovascular comorbidities was used to group patients in three groups: cardiovascular controls, subjects with a cardiovascular condition and subjects who suffered from a cardiovascular event.

In Chapter 11, several pulse oximetry markers were investigated to differentiate between these groups and an LS-SVM classifier was trained to separate CV event patients from CV controls. The findings of this study suggest that SpO_2-based features can be of added value to assess the CV status of OSA patients. More severe oxygen desaturations and incomplete resaturations to the baseline SpO_2 values were observed for patients with CV comorbidities. Moreover, a PPG feature pointing to less arousals in these patients was obtained. The investigated PPG-based features, however, did not improve the LS-SVM classification performance.

The obtained SpO_2-based markers were further studied in Chapter 11. Additional statistical tests were performed comparing the different CV conditions and the influence of drug use. Moreover, a multi-level ICS model was trained to assess the CV risk in OSA patients. This model included the age, BMI and the SpO_2 parameters and obtained good classification performances on the UZ Leuven data sets. The model was also tested on a set of healthy volunteers, which were all classified as having no CV risk. Moreover, a subset of the SHHS was included in order to assess the predictive performance of the extracted SpO_2 parameters. The SpO_2 trends observed in the UZ Leuven data sets, however, could not be confirmed in this data set. Several explanations were given why the SHHS is not optimal for this application. Clinical OSA data sets including a long-term cardiovascular follow-up should be used to properly validate the predictive power of these SpO_2 markers.

13.1.3 Algorithmic contributions

In Chapter 4, the extension of the binary ICS model to multilevel problems was proposed. Moreover, monotonicity constraints were introduced in order to stabilize the ICS scores when only a small set of training samples is available. These extensions to the ICS model can also be useful in other applications and were added to the ICS toolbox, publicly available at [49].

Improvements to the random forest based feature selection were proposed in Chapter 5. By selecting the OOB samples per patient, a patient independent validation of the feature selection is ensured. Moreover, using the Cohen's κ value as performance estimate leads to better results in unbalanced and multi-class problems. Furthermore, the inclusion of the feature correlation in the predictor importance definition, avoids the selection of redundant features and results in a more pruned feature set. This novel feature selection method is generally applicable when a reduced feature set is needed to enhance the interpretability of the model.

13.2 Future directions

In this PhD important steps were taken towards the development of multimodal OSA screening methods and the cardiovascular phenotyping of OSA patients. Nevertheless, additional research is needed in order to further improve and validate the proposed methods. The next sections will give an overview of possible future research steps.

13.2.1 Multimodal OSA screening methods

The SpO_2-based OSA screening algorithm obtained very good performances on the UZ Leuven data sets. Nevertheless, some patients with primarily hypopneas accompanied by arousals without any oxygen desaturation, will be missed. The clinical relevance of these OSA cases should be investigated and alternative screening methods should be proposed to detect them. In order to further validate the SpO_2-based screening algorithm, a new clinical study is planned. Within this study, patients will undergo an overnight PSG at the hospital, followed by 7 nights of oximetry recordings at home. Using this data it will be possible to assess the night-to-night variability of the SpO_2-based AHI estimate and to compare the results from the in-home monitoring to the in-hospital PSG. The oximetry measurements will be performed using a transmissive finger clip sensor. Reflective oximetry sensors might be more comfortable for the patients

for long-term monitoring, but extracting the oxygen saturation levels is more challenging in reflective sensors. Studies have shown that only on the finger and the forehead, reliable SpO_2 recordings could be made using a reflective sensor [80]. Therefore, at the moment it is not possible to reliably record the SpO_2 signal from watch-like devices, which are popular for long-term personal health tracking. A boost in the development and commercialization of SpO_2-based OSA screening methods is expected when technological advancements enable reliable SpO_2 measurements from these devices. The SpO_2-based OSA screening method developed within this PhD can be very useful for wearable oximetry devices, as it is based on easy-to-compute parameters, and obtains very good classification performances.

The proposed OSA detection methods were all based on manually defined features using previous knowledge of the effect apneas have on the signals. More and more researchers have, however, moved towards representation learning (i.e., automatic feature extraction) using deep learning. These methods can be useful for OSA detection methods, as in this case, the interpretability of the model is less crucial. In the master thesis of Borys Berlog [15], we have studied the use of CNN models for the detection of apneic events in SpO_2 and ECG signals of the UZ Leuven OSA+ data set. Based on the SpO_2 signal, the developed CNN could not outperform the feature-based method presented in Chapter 6. This might be due to the relatively simple structure of the SpO_2 signals from which the relevant information can be extracted using simple features. The ECG, on the other hand, is a more complex signal from which the deep learning methods might gather extra information. The ECG CNN model developed in this master thesis obtained an AUC of 75.5 %, which is slightly higher than the AUC of 73.5 % obtained using the feature-based model. These results show the potential of deep learning methods for OSA detection based on ECG signals. Using more complex architectures or using recurrent neural networks and including larger data sets might further improve these results. These methods might also be useful for the improvement of the PPG-based apnea detection algorithms.

In order to improve the development of ECG and PPG-based OSA detection algorithms, large publicly available data sets are needed, which can be used as a standard to compare methods. For ECG-based methods, the Physionet Apnea-ECG [104] data set is often used. But, this data set is rather small, especially when deep learning methods are considered. Moreover, the data set is not challenging enough, making it difficult to differentiate between methods based on this data set.

Different ways of combining the unimodal apnea detection methods were investigated in this thesis, but none of them could improve the SpO_2-based OSA detection algorithm. Moreover, in the master thesis of Borys Berlog [15] the merging of the SpO_2 and ECG CNNs also did not improve the SpO_2-based

classification performance. “Smarter” ways of combining the different signal modalities should thus be considered. This might, for example, consist of a sequential combination of classifiers, in which the ECG or PPG signals are only considered when no large oxygen desaturations could be observed in the SpO_2 signal. Another option would be to define multimodal features which study the interactions between the SpO_2 and ECG or PPG signals. Or to investigate methods which can take into account the delays between the effect of the apneic event in the different signals, and which can exploit sequential information. Although, in our experiment using the 4-state hidden Markov model the inclusion of sequential information did not improve the performance.

The performance results on the experiments using ccECG, ccBioz and wearable PPG signals were promising, but further research is needed to improve the results. To improve the apnea detection using ccECG signals, more research needs to be done into the effect that respiration has on these signals in order to obtain better EDR estimates. Moreover, larger data sets should be collected, such that the models could be retrained based on the non-contact data. Using the ccBioz signals, good performances were obtained, but this signal was only available for a small subset of the data. These results should thus be further validated on larger data sets. Therefore, additional measurements in the sleep lab including these capacitively-coupled sensors are planned.

In order to improve the apnea detection performance using the wearable PPG recordings from the Byteflies sensors, better artefact detection algorithms should be developed. Moreover, it should be investigated whether the information from the different light frequencies could be combined in order to obtain cleaner signals. Additionally, the parameter selection should be optimized for the wearable data.

13.2.2 Characterization of OSA patients beyond the AHI

The most important point to improve and further validate the SpO_2-based markers for the cardiovascular assessment of OSA patients is to obtain extra data sets. Ideally, a clinical data set recorded on subjects suspected of having OSA which includes a long-term cardiovascular follow-up should be gathered. On such a data set it can be tested whether the SpO_2-based markers could be used to predict the development or worsening of cardiovascular diseases. The population should be younger than those included in the SHHS, as the older the people get, the higher the chance that they have developed cardiovascular problems which were unrelated to the OSA. Moreover, the data set should be expanded such that the full range of patient characteristics is covered by the different cardiovascular classes. The UZ Leuven data sets, which are used at

the moment, lack older cardiovascular controls with a high BMI and younger cardiovascular event subjects with a low BMI. Therefore, the model tends to rely heavily on the age and BMI to predict the cardiovascular status of patients.

The studied PPG features did not turn out to have an added value to our model to assess the CV status in OSA subjects. The addition of novel PPG features could be investigated, as well as new artefact detection techniques to improve the reliability of the PPG features.

Instead of looking to the diagnosed cardiovascular diseases in OSA subjects, it could also be interesting to study the relationship between the SpO_2-based markers and characteristics in the ECG signals used to assess the cardiovascular status. Heart rate variability parameters or the occurrence of ectopic beats could for example be used. In Chapter 12, for example, a worsening of the SpO_2-based markers could be observed for 7 control patients which had a large number of ectopic beats in their ECG signals. It might also be interesting to investigate whether the occurrence of ectopic beats in these patients is linked to certain sleep stages or to the occurrence of apneas, oxygen desaturations or arousals.

The SpO_2-based markers for the cardiovascular assessment of OSA patients will also be studied in the planned clinical study which will include 7 nights of oximetry recordings. As such, the night-to-night variability of these markers will be assessed. This variability should be as low as possible, as the cardiovascular risk of patients should not differ.

Alternative ways of phenotyping OSA patients, next to the cardiovascular comorbidities, have been proposed in the literature. Some examples are the clustering based on clinical presentation, risk factors and comorbidities [8], and the differentiation based on physiological traits causing OSA [44]. In order to obtain a full picture to phenotype OSA patients, these different methods should be combined and connections should be sought. The proposed SpO_2-based markers might for example be linked to a specific physiological trait proposed in [44].

The further validation of the novel SpO_2 markers on prospective clinical sets, in correlation with changes in the ECG might potentially lead to a better understanding of the physiological effect of apneas on the cardiovascular system. In combination with alternative OSA traits, this might in the future open the door to new ways of phenotyping OSA patients in clinical practice beyond the AHI.

Appendix A

SpO_2 based apnea detection GUI

In this PhD, a GUI was developed in Python using the Tkinter library to estimate the AHI and mICS cardiac risk score parameters from the SpO_2 signal. A screenshot of the GUI is presented in Figure A.1.

The user should first select the input data folder. The .mat and .EDF files contained in this folder will then show up in the listbox on the left. The user can then select the files that need to be processed. Checkboxes are available to select all .mat and .EDF files within the folder, or only the .mat files.

When the signal files are given in the EDF format, the user can extract the SpO_2 signals from these files using the "Extract SpO_2 from EDF" button. The program will then search for the "SaO2" channel in the EDF files, and will save these as separate .mat files with the same name as the original .EDF files.

Next, the SpO_2 parameters can be extracted from the selected input files by clicking the "Compute SpO_2 parameters" button. This function takes as input the sampling frequency, age and BMI values. The sampling frequency is by default 500 Hz. When a value lower than 1 is entered, an error will be given. No default values for the Age and BMI are included. A warning will be given when these fields are not correctly filled out, but the user can choose to proceed without providing these values. In this case, the mICS risk score will not be reliable, but the AHI estimates and SpO_2 parameters needed to estimate the risk score will be computed. When a large number of recordings from different patients needs to be processed, this approach might be preferred as the age and

BMI should be adapted for each one of them. Based on the extracted SpO_2 parameters, the age and BMI, the mICS risk can easily be computed afterwards. When the computations are completed, the computed parameters will be shown in the results table on the top right of the GUI. These include the AHI estimate, predicted OSA severity level, mICS risk score, and the 3 SpO_2 parameters to compute this risk score. The computation of these parameters is also possible directly on the EDF files, the intermediate .mat files will then not be saved.

The estimated AHI values can be plotted using the "Plot AHIs" button. Moreover, the SpO_2 signals can be visually checked using the "Plot SpO_2" button, as can be seen from the example in Figure A.2. For this, only one input signal can be selected at a time from the listbox. An error will be given in case the user tries to plot multiple signals. A plotting toolbar is available to zoom in and out and to save the plots.

The obtained results given in the results table can finally be saved as a .csv file using the "Save results" button.

On the lower left of the GUI, a feedback window is provided, which will give status updates on the ongoing processes.

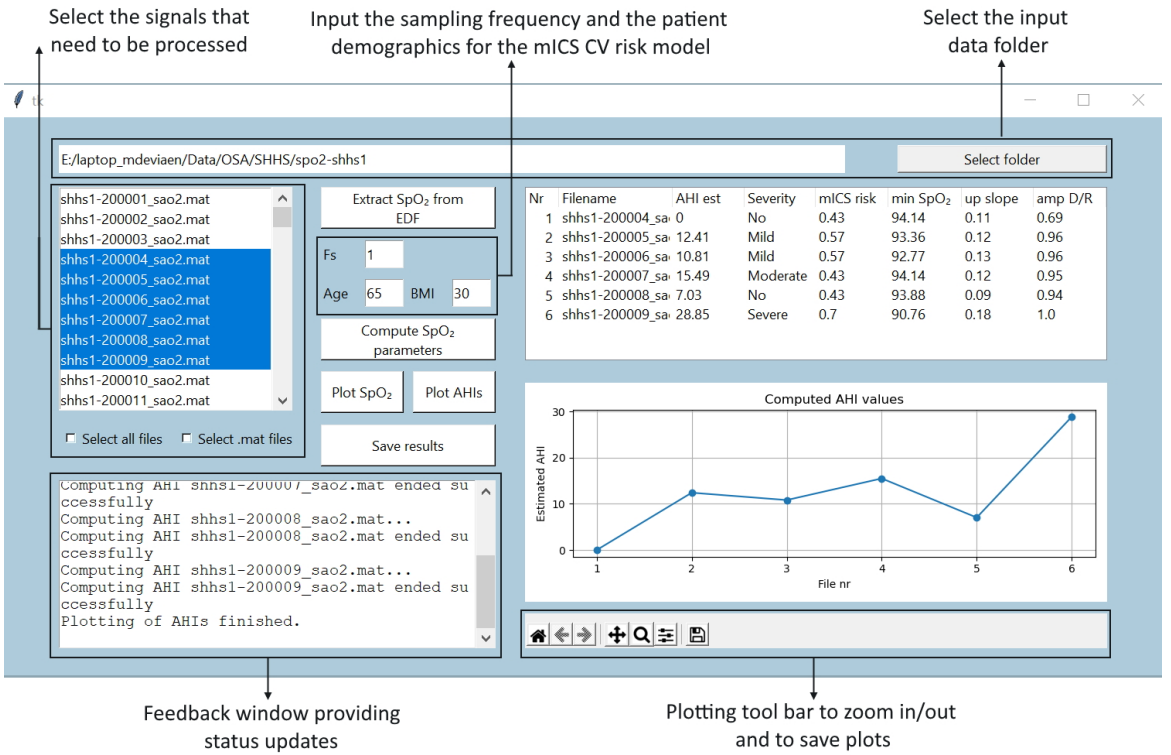


Figure A.1: Graphical user interface to estimate the AHI and mICS cardiac risk score parameters from the SpO_2.

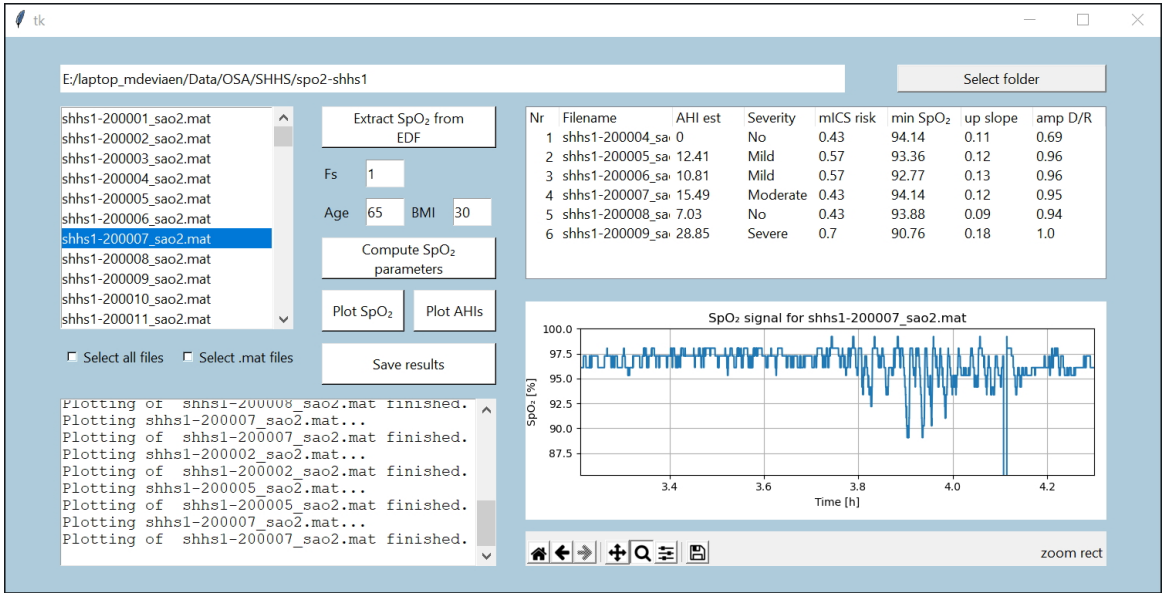


Figure A.2: Example of SpO_2 plotting in the GUI.

Bibliography

[1] AASM Task Force. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. *Sleep 22*, 5 (1999), 667–689.

[2] Al-Angari, H. M., and Sahakian, A. V. Use of sample entropy approach to study heart rate variability in obstructive sleep apnea syndrome. *IEEE Transactions on Biomedical Engineering 54*, 10 (2007), 1900–1904.

[3] Allen, J. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement 28*, 3 (2007), R1.

[4] Alvarez, D., Hornero, R., Marcos, J. V., Wessel, N., Penzel, T., Glos, M., and Del Campo, F. Assessment of feature selection and classification approaches to enhance information from overnight oximetry in the context of apnea diagnosis. *International journal of neural systems 23*, 05 (2013), 1350020.

[5] Alvarez et al, D. Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. *IEEE Transactions on Biomedical Engineering 57*, 12 (2010), 2816–2824.

[6] Apple Inc. Beddit sleep monitor. https://www.beddit.com/, 2019.

[7] Azarbarzin et al, A. The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the osteoporotic fractures in men study and the sleep heart health study. *European heart journal* (2018).

[8] Bailly et al, S. Obstructive sleep apnea: a cluster analysis at time of diagnosis. *PLoS One 11*, 6 (2016), e0157318.

[9] Barbieri, R., Matten, E. C., Alabi, A. A., and Brown, E. N. A point-process model of human heartbeat intervals: new definitions of heart

rate and heart rate variability. *American Journal of Physiology-Heart and Circulatory Physiology 288*, 1 (2005), H424–H435.

[10] Bauer et al, A. Phase-rectified signal averaging detects quasi-periodicities in non-stationary data. *Physica A: Statistical Mechanics and its Applications 364* (2006), 423–434.

[11] Behar, J., Roebuck, A., Shahid, M., Daly, J., Hallack, A., Palmius, N., Stradling, J., and Clifford, G. D. Sleepap: an automated obstructive sleep apnoea screening application for smartphones. *IEEE journal of biomedical and health informatics 19*, 1 (2014), 325–331.

[12] Behar, J. A., Palmius, N., Li, Q., Garbuio, S., Rizzatti, F. P., Bittencourt, L., Tufik, S., and Clifford, G. D. Feasibility of single channel oximetry for mass screening of obstructive sleep apnea. *EClinicalMedicine 11* (2019), 81–88.

[13] Ben-David, A. Comparison of classification accuracy using cohen's weighted kappa. *Expert Systems with Applications 34*, 2 (2008), 825–832.

[14] Benjafield, A. V., Ayas, N. T., Eastwood, P. R., Heinzer, R., Ip, M. S., Morrell, M. J., Nunez, C. M., Patel, S. R., Penzel, T., Pépin, J.-L., et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine 7*, 8 (2019), 687–698.

[15] Berlog, B. Developing deep neural network for sleep apnea detection. *KU Leuven* (2019).

[16] Berntson, G. G., Thomas Bigger Jr, J., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., et al. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology 34*, 6 (1997), 623–648.

[17] Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., Marcus, C. L., Mehra, R., Parthasarathy, S., Quan, S. F., et al. Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events. *Journal of clinical sleep medicine 8*, 05 (2012), 597–619.

[18] Billiet et al, L. Interval coded scoring: a toolbox for interpretable scoring systems. *PeerJ Computer Science 4* (2018), e150.

[19] Bradley, T. D., and Floras, J. S. Obstructive sleep apnoea and its cardiovascular consequences. *The Lancet 373*, 9657 (2009), 82–93.

[20] Brager, C. A., and Kryger, M. A new approach: Integrating sleep science and consumer wearables to improve access and drive greater engagement.

[21] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[22] Bsoul, M., Minn, H., and Tamil, L. Apnea medassist: real-time sleep apnea monitor using single-lead ecg. *IEEE Transactions on Information Technology in Biomedicine 15*, 3 (2010), 416–427.

[23] Burgos et al, A. Real-time detection of apneas on a pda. *IEEE Transactions on Information Technology in Biomedicine 14*, 4 (2010), 995–1002.

[24] Castro, I., Varon, C., Moeyersons, J., Gomez, A. V., Morales, J., Deviaene, M., Torfs, T., Van Huffel, S., Puers, R., and Van Hoof, C. Data quality assessment of capacitively-coupled ecg signals. In *2019 Computing in Cardiology (CinC)* (2019), IEEE, pp. Page–1.

[25] Castro, I., Varon, C., Torfs, T., Van Huffel, S., Puers, R., and Van Hoof, C. Evaluation of a multichannel non-contact ecg system and signal quality algorithms for sleep apnea detection and monitoring. *Sensors 18*, 2 (2018), 577.

[26] Castro, I. D., Patel, A., Torfs, T., Puers, R., and Van Hoof, C. Capacitive multi-electrode array with real-time electrode selection for unobtrusive ecg & bioz monitoring. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2019), IEEE, pp. 5621–5624.

[27] Castro Miller, I. Robust capacitively-coupled electrocardiography and bioimpedance for unobtrusive health monitoring.

[28] Chiner et al, E. Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies? *Thorax 54*, 11 (1999), 968–971.

[29] Ciołek, M., Niedźwiecki, M., Sieklicki, S., Drozdowski, J., and Siebert, J. Automated detection of sleep apnea and hypopnea events based on robust airflow envelope tracking in the presence of breathing artifacts. *IEEE journal of biomedical and health informatics 19*, 2 (2014), 418–429.

[30] Collop, N. A., Tracy, S. L., Kapur, V., Mehra, R., Kuhlmann, D., Fleishman, S. A., and Ojile, J. M. Obstructive sleep apnea

devices for out-of-center (ooc) testing: technology evaluation. *Journal of Clinical Sleep Medicine 7*, 05 (2011), 531–548.

[31] Cover, T., and Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory 13*, 1 (1967), 21–27.

[32] D'agostino et al, R. B. General cardiovascular risk profile for use in primary care. *Circulation 117*, 6 (2008), 743–753.

[33] De Chazal, P., Heneghan, C., Sheridan, E., Reilly, R., Nolan, P., and O'Malley, M. Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea. *IEEE Transactions on Biomedical Engineering 50*, 6 (2003), 686–696.

[34] Dean, D. A., Goldberger, A. L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., Sahoo, S. S., Jayapandian, C. P., Cui, L., Morrical, M. G., et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep 39*, 5 (2016), 1151–1164.

[35] Deviaene, M., , Huysmans, D., Castro, I., Borzée, P., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Sleep monitoring in adults using wearables and unobtrusive technology. In *Wearable/Personal monitoring devices present to future*, G. Gargiulo and G. R. Naik, Eds. Springer, 2020, pp. –.

[36] Deviaene, M., Borzée, P., Buyse, B., Testelmans, D., Van Huffel, S., and Varon, C. Pulse oximetry markers for cardiovascular disease in sleep apnea. In *2019 Computing in Cardiology (CinC)* (2019), IEEE, pp. Page–1.

[37] Deviaene, M., Borzée, P., Van Gilst, M., Van Dijk, J., Overeem, S., Buyse, B., Testelmans, D., Van Huffel, S., and Varon, C. Multilevel interval coded scoring to assess the cardiovascular status of sleep apnea patients using oxygen saturation markers. *IEEE Transactions on Biomedical Engineering 67*, 10 (2020), 2839–2848.

[38] Deviaene, M., Castro, I. D., Borzée, P., Patel, A., Torfs, T., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Capacitively-coupled ecg and respiration for the detection of sleep apnea. *Submitted to Physiological Measurement* (2020).

[39] Deviaene, M., Lázaro, J., Huysmans, D., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Sleep apnea detection using pulse photoplethysmography. In *2018 Computing in Cardiology Conference (CinC)* (2018), vol. 45, IEEE, pp. 1–4.

[40] Deviaene, M., Testelmans, D., Borzée, P., Buyse, B., Van Huffel, S., and Varon, C. Feature selection algorithm based on random forest applied to sleep apnea detection. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2019), IEEE, pp. 2580–2583.

[41] Deviaene, M., Testelmans, D., Buyse, B., Borzée, P., Van Huffel, S., and Varon, C. Automatic screening of sleep apnea patients based on the spo_2 signal. *IEEE journal of biomedical and health informatics 23*, 2 (2018), 607–617.

[42] Deviaene, M., Varon, C., Testelmans, D., Buyse, B., and Van Huffel, S. Assessing cardiovascular comorbidities in sleep apnea patients using spo 2. In *2017 Computing in Cardiology (CinC)* (2017), IEEE, pp. 1–4.

[43] Devos, A. K., Van Huffel, S., Simonetti, A. W., van der Graaf, M., Heerschap, A., and Buydens, L. M. Classification of brain tumours by pattern recognition of magnetic resonance imaging and spectroscopic data. In *Outcome Prediction in Cancer*. Elsevier, 2007, pp. 285–318.

[44] Eckert, D. J. Phenotypic approaches to obstructive sleep apnoea–new pathways for targeted therapy. *Sleep medicine reviews 37* (2018), 45–59.

[45] Eckert, D. J., and Malhotra, A. Pathophysiology of adult obstructive sleep apnea. *Proceedings of the American thoracic society 5*, 2 (2008), 144–153.

[46] Eckert, D. J., and Younes, M. K. Arousal from sleep: implications for obstructive sleep apnea pathogenesis and treatment. *Journal of Applied Physiology 116*, 3 (2014), 302–313.

[47] Emfit. Emfit qs+active sleep and recovery monitor for the active life. https://www.emfit.com/active-life-sleep-and-recovery, 2020.

[48] Endeshaw et al, Y. Snoring, daytime sleepiness, and incident cardiovascular disease in the health, aging, and body composition study. *Sleep 36*, 11 (2013), 1737–1745.

[49] ESAT-STADIUS. Interval coded scoring toolbox. https://www.esat.kuleuven.be/stadius/icstoolbox/.

[50] Faust, O., Acharya, U. R., Ng, E., and Fujita, H. A review of ecg-based diagnosis support systems for obstructive sleep apnea. *Journal of Mechanics in Medicine and Biology 16*, 01 (2016), 1640004.

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

[52] Fonseca, P., Weysen, T., Goelema, M. S., Møst, E. I., Radha, M., Lunsingh Scheurleer, C., van den Heuvel, L., and Aarts, R. M. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep 40*, 7 (2017), zsx097.

[53] Gil, E., Vergara, J. M., and Laguna, P. Detection of decreases in the amplitude fluctuation of pulse photoplethysmography signal as indication of obstructive sleep apnea syndrome in children. *Biomedical Signal Processing and Control 3*, 3 (2008), 267–277.

[54] Goldberger et al, A. L. Physiobank, physiotoolkit, and physionet. *Circulation 101*, 23 (2000), e215–e220.

[55] Gottlieb et al, D. J. Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure the sleep heart health study. *Circulation 122*, 4 (2010), 352–360.

[56] Gottlieb et al, J. D. Hypoxia, not the frequency of sleep apnea, induces acute hemodynamic stress in patients with chronic heart failure. *Journal of the American College of Cardiology 54*, 18 (2009), 1706–1712.

[57] Guilleminault, C., Winkle, R., Connolly, S., Melvin, K., and Tilkian, A. Cyclical variation of the heart rate in sleep apnoea syndrome: Mechanisms, and usefulness of 24 h electrocardiography as a screening technique. *The Lancet 323*, 8369 (1984), 126–131.

[58] Guyon, I., and Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research 3*, Mar (2003), 1157–1182.

[59] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[60] Heinzer et al, R. Prevalence of sleep-disordered breathing in the general population: the hypnolaus study. *The Lancet Respiratory Medicine 3*, 4 (2015), 310–318.

[61] Holland, P. W., and Welsch, R. E. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods 6*, 9 (1977), 813–827.

[62] Huysmans, D., Borzée, P., Testelmans, D., Buyse, B., Willemen, T., Huffel, S. V., and Varon, C. Evaluation of a commercial ballistocardiography sensor for sleep apnea screening and sleep monitoring. *Sensors 19*, 9 (2019), 2133.

[63] Hwang, S. H., Lee, Y. J., Jeong, D.-U., Park, K. S., et al. Oxygen desaturation index estimation through unconstrained cardiac sympathetic activity assessment using three ballistocardiographic systems. *Respiration 92*, 2 (2016), 90–97.

[64] Iber et al, C. The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications. *American Academy of Sleep Medicine* (2007).

[65] Itamar medical. Watchpat. https://www.itamar-medical.com/, 2020.

[66] Johns, M. W. A new method for measuring daytime sleepiness: the epworth sleepiness scale. *sleep 14*, 6 (1991), 540–545.

[67] Kagawa, M., Tojima, H., and Matsui, T. Non-contact diagnostic system for sleep apnea–hypopnea syndrome based on amplitude and phase analysis of thoracic and abdominal doppler radars. *Medical & biological engineering & computing 54*, 5 (2016), 789–798.

[68] Kantelhardt et al, J. W. Phase-rectified signal averaging for the detection of quasi-periodicities and the prediction of cardiovascular risk. *Chaos: An Interdisciplinary Journal of Nonlinear Science 17*, 1 (2007), 015112.

[69] Kemp, B., Janssen, A. J., and Van Der Kamp, B. Body position can be monitored in 3d using miniature accelerometers and earth-magnetic field sensors. *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control 109*, 6 (1998), 484–488.

[70] Kendzerska et al, T. Obstructive sleep apnea and risk of cardiovascular events and all-cause mortality: a decade-long historical cohort study. *PLoS medicine 11*, 2 (2014), e1001599.

[71] Koley, B. L., and Dey, D. Automatic detection of sleep apnea and hypopnea events from single channel measurement of respiration signal employing ensemble binary svm classifiers. *Measurement 46*, 7 (2013), 2082–2092.

[72] Koley, B. L., and Dey, D. On-line detection of apnea/hypopnea events using spo_2 signal: A rule-based approach employing binary classifier models. *IEEE journal of biomedical and health informatics 18*, 1 (2013), 231–239.

[73] Kontaxis, S., Lázaro, J., Corino, V. D., Sandberg, F., Bailón, R., Laguna, P., and Sörnmo, L. Ecg-derived respiratory rate in atrial fibrillation. *IEEE Transactions on Biomedical Engineering 67*, 3 (2019), 905–914.

[74] Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering 160*, 1 (2007), 3–24.

[75] Krieger et al, J. Respiratory effort during obstructive sleep apnea: role of age and sleep state. *Chest 112*, 4 (1997), 875–884.

[76] Kulkas, A., Tiihonen, P., Julkunen, P., Mervaala, E., and Töyräs, J. Novel parameters indicate significant differences in severity of obstructive sleep apnea with patients having similar apnea–hypopnea index. *Medical & biological engineering & computing 51*, 6 (2013), 697–708.

[77] Landolt, H.-P., Holst, S. C., Sousek, A., Bassetti, C., Dogas, Z., and Peigneux, P. Effects of acute and chronic sleep deprivation. In *Sleep medicine textbook*, C. Bassetti, Z. Dogas, and P. Peigneux, Eds. European Sleep Research Society, 2014, ch. A-5, pp. 49–61.

[78] Lázaro, J., Gil, E., Deviaene, M., Bailón, R., Testelmans, D., Buyse, B., Varón, C., and Van Huffel, S. Pulse photoplethysmography derived respiration for obstructive sleep apnea detection. In *2017 Computing in Cardiology (CinC)* (2017), IEEE, pp. 1–4.

[79] Lázaro, J., Gil, E., Vergara, J. M., and Laguna, P. Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children. *IEEE journal of biomedical and health informatics 18*, 1 (2013), 240–246.

[80] Longmore, S. K., Lui, G. Y., Naik, G., Breen, P. P., Jalaludin, B., and Gargiulo, G. D. A comparison of reflective photoplethysmography for detection of heart rate, blood oxygen saturation, and respiration rate at various anatomical locations. *Sensors 19*, 8 (2019), 1874.

[81] Mador, M. J., Kufel, T. J., Magalang, U. J., Rajesh, S., Watwe, V., and Grant, B. J. Prevalence of positional sleep apnea in patients undergoing polysomnography. *Chest 128*, 4 (2005), 2130–2137.

[82] Malik, M. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology. *Annals of Noninvasive Electrocardiology 1*, 2 (1996), 151–181.

[83] Marin, J. M., Carrizo, S. J., Vicente, E., and Agusti, A. G. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. *The Lancet 365*, 9464 (2005), 1046–1053.

[84] Medatec. Dreamscan 2. https://www.medatec.eu/index.php/nl/slaap, 2020.

[85] Mendonça, F., Mostafa, S. S., Ravelo-García, A. G., Morgado-Dias, F., and Penzel, T. Devices for home detection of obstructive sleep apnea: A review. *Sleep medicine reviews 41* (2018), 149–160.

[86] Mendonça, F., Mostafa, S. S., Ravelo-García, A. G., Morgado-Dias, F., and Penzel, T. A review of obstructive sleep apnea detection approaches. *IEEE journal of biomedical and health informatics 23*, 2 (2018), 825–837.

[87] Messineo, L., Magri, R., Corda, L., Pini, L., Taranto-Montemurro, L., and Tantucci, C. Phenotyping-based treatment improves obstructive sleep apnea symptoms and severity: a pilot study. *Sleep and Breathing 21*, 4 (2017), 861–868.

[88] Meyer, D., Leisch, F., and Hornik, K. The support vector machine under test. *Neurocomputing 55*, 1-2 (2003), 169–186.

[89] Milagro, J., Deviaene, M., Gil, E., Lázaro, J., Buyse, B., Testelmans, D., Borzée, P., Willems, R., Van Huffel, S., Bailón, R., et al. Autonomic dysfunction increases cardiovascular risk in the presence of sleep apnea. *Frontiers in Physiology 10* (2019), 620.

[90] Moeyersons, J., Smets, E., Morales, J., Villa, A., De Raedt, W., Testelmans, D., Buyse, B., Van Hoof, C., Willems, R., Van Huffel, S., et al. Artefact detection and quality assessment of ambulatory ecg signals. *Computer methods and programs in biomedicine 182* (2019), 105050.

[91] Moody, G., Mark, R., Goldberger, A., and Penzel, T. Stimulating rapid research advances via focused competition: The computers in cardiology challenge 2000. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)* (2000), IEEE, pp. 207–210.

[92] Moody, G. B., Mark, R. G., Zoccola, A., and Mantero, S. Derivation of respiratory signals from multi-lead ecgs. *Computers in cardiology 12*, 1985 (1985), 113–116.

[93] Morales, J. F., Varon, C., Deviaene, M., Borzée, P., Testelmans, D., Buyse, B., and Van Huffel, S. Sleep apnea hypopnea syndrome classification in spo 2 signals using wavelet decomposition and phase space reconstruction. In *2017 IEEE 14th international conference on wearable and implantable body sensor networks (BSN)* (2017), IEEE, pp. 43–46.

[94] Morales Tellez, J. F., Deviaene, M., Milagro, J., Testelmans, D., Buyse, B., Willems, R., Orini, M., Van Huffel, S., Bailón, R., and Varon, C. Evaluation of methods to characterize the change of the respiratory sinus arrhythmia with age in sleep apnea patients. In *Proceedings of EMBC* (2019), IEEE.

[95] Moret-Bonillo, V., Alvarez-Estévez, D., Fernández-Leal, A., and Hernández-Pereira, E. Intelligent approach for analysis of respiratory signals and oxygen saturation in the sleep apnea/hypopnea syndrome. *The open medical informatics journal 8* (2014), 1.

[96] Morgenthaler, T., Alessi, C., Friedman, L., Owens, J., Kapur, V., Boehlecke, B., Brown, T., Chesson Jr, A., Coleman, J., Lee-Chiong, T., et al. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. *Sleep 30*, 4 (2007), 519–529.

[97] Mostafa, S. S., Mendonça, F., Ravelo-García, A. G., and Morgado-Dias, F. A systematic review of detecting sleep apnea using deep learning. *Sensors 19*, 22 (2019), 4934.

[98] Nonin. Wristox. https://www.nonin.com/products/3150-usb/, 2020.

[99] O. M. L. Center. Tabulated molar extinction coefficient for hemoglobin in water. https://omlc.org/spectra/hemoglobin/summary.html.

[100] Oksenberg, A., Khamaysi, I., Silverberg, D. S., and Tarasiuk, A. Association of body position with severity of apneic events in patients with severe nonpositional obstructive sleep apnea. *Chest 118*, 4 (2000), 1018–1024.

[101] Papini, G. B., Fonseca, P., Margarito, J., van Gilst, M. M., Overeem, S., Bergmans, J. W., and Vullings, R. On the generalizability of ecg-based obstructive sleep apnea monitoring: merits and limitations of the apnea-ecg database. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018), IEEE, pp. 6022–6025.

[102] Papini, G. B., Fonseca, P., van Gilst, M. M., Bergmans, J. W., Vullings, R., and Overeem, S. Respiratory activity extracted from wrist-worn reflective photoplethysmography in a sleep-disordered population. *Physiological Measurement* (2020).

[103] Peng et al, H. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence 27*, 8 (2005), 1226–1238.

[104] Penzel, T., Moody, G. B., Mark, R. G., Goldberger, A. L., and Peter, J. H. The apnea-ecg database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)* (2000), IEEE, pp. 255–258.

[105] Peppard et al, P. E. The impact of obesity on oxygen desaturation during sleep-disordered breathing. *American journal of respiratory and critical care medicine 180*, 8 (2009), 788–793.

[106] Pevernagie, D., Sastry, M., and Vanderveken, O. Treatment of respiratory sleep disorders. In *Sleep medicine textbook*, C. Bassetti, Z. Dogas, and P. Peigneux, Eds. European Sleep Research Society, 2014, ch. D-5, pp. 259–274.

[107] Quan, S., Gillin, J. C., Littner, M., and Shepard, J. Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. editorials. *Sleep (New York, NY) 22*, 5 (1999), 662–689.

[108] Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., et al. The sleep heart health study: design, rationale, and methods. *Sleep 20*, 12 (1997), 1077–1085.

[109] Rabiner, L., and Juang, B. An introduction to hidden markov models. *ieee assp magazine 3*, 1 (1986), 4–16.

[110] Rangayyan, R. M. *Biomedical signal analysis*, vol. 33. John Wiley & Sons, 2015.

[111] Redline et al, S. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. *Sleep 21*, 7 (1998), 759–767.

[112] Respironics. Alice nightone. https://www.usa.philips.com/healthcare/product/HC1114801/alice-nightone-home-sleep-testing-device, 2020.

[113] Rolón et al, R. Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection. *Biomedical Signal Processing and Control 33* (2017), 358–367.

[114] Sánchez-de-la Torre, M., Campos-Rodriguez, F., and Barbé, F. Obstructive sleep apnoea and cardiovascular disease. *The Lancet Respiratory Medicine 1*, 1 (2013), 61–72.

[115] Sateia, M. J. International classification of sleep disorders. *Chest 146*, 5 (2014), 1387–1394.

[116] Schlotthauer et al, G. Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry. *Medical engineering & physics 36*, 8 (2014), 1074–1080.

[117] Schnall, R. P., Shlitner, A., Sheffy, J., Kedar, R., and Lavie, P. Periodic, profound peripheral vasoconstriction as a new marker of obstructive sleep apnea. *Sleep 22*, 7 (1999), 939–946.

[118] Sharma et al, A. M. High prevalence and poor control of hypertension in primary care: cross-sectional study. *Journal of hypertension 22*, 3 (2004), 479–486.

[119] Solà-Soler, J., Fiz, J. A., Morera, J., and Jané, R. Multiclass classification of subjects with sleep apnoea–hypopnoea syndrome through snoring analysis. *Medical engineering & physics 34*, 9 (2012), 1213–1220.

[120] Sommermeyer et al, D. The use of overnight pulse wave analysis for recognition of cardiovascular risk factors and risk: a multicentric evaluation. *Journal of hypertension 32*, 2 (2014), 276–285.

[121] Stradling, J., and Davies, R. Sleep· 1: Obstructive sleep apnoea/hypopnoea syndrome: definitions, epidemiology, and natural history. *Thorax 59*, 1 (2004), 73–78.

[122] Strobl et al, C. Conditional variable importance for random forests. *BMC bioinformatics 9*, 1 (2008), 307.

[123] Suykens, J. A., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., Suykens, J., and Van Gestel, T. *Least squares support vector machines*, vol. 4. World Scientific, 2002.

[124] Tamura, T., Maeda, Y., Sekine, M., and Yoshida, M. Wearable photoplethysmographic sensors—past and present. *Electronics 3*, 2 (2014), 282–302.

[125] Tang, J., Alelyani, S., and Liu, H. Feature selection for classification: A review. *Data classification: Algorithms and applications* (2014), 37.

[126] Terrill, P. I. A review of approaches for analysing obstructive sleep apnoea-related patterns in pulse oximetry data. *Respirology* (2019).

[127] Tremper, K. K., and Barker, S. J. Pulse oximetry. *Anesthesiology: The Journal of the American Society of Anesthesiologists 70*, 1 (1989), 98–108.

[128] Uddin, M., Chow, C., and Su, S. Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: a systematic review. *Physiological measurement 39*, 3 (2018), 03TR01.

[129] Varon, C. Mining the ecg: algorithms and applications.

[130] Varon, C., Jansen, K., Lagae, L., and Van Huffel, S. Can ecg monitoring identify seizures? *Journal of electrocardiology 48*, 6 (2015), 1069–1074.

[131] Varon, C., Lázaro, J., Bolea, J., Hernando, A., Aguiló, J., Gil, E., Van Huffel, S., and Bailón, R. Unconstrained estimation of hrv indices after removing respiratory influences from heart rate. *IEEE journal of biomedical and health informatics* (2018).

[132] Varon, C., Morales, J., Lázaro, J., Orini, M., Deviaene, M., Kontaxis, S., Testelmans, D., Buyse, B., Borzée, P., Sörnmo, L., et al. A comparative study of ecg-derived respiration in ambulatory monitoring using the single-lead ecg. *Scientific Reports 10*, 1 (2020), 1–14.

[133] Varon, C., and Van Huffel, S. Complexity and nonlinearities in cardiorespiratory signals in sleep and sleep apnea. In *Complexity and Nonlinearity in Cardiovascular Signals*. Springer, 2017, pp. 503–537.

[134] Varon et al, C. Noise level estimation for model selection in kernel pca denoising. *IEEE transactions on neural networks and learning systems 26*, 11 (2015), 2650–2663.

[135] Varon et al, C. A novel algorithm for the automatic detection of sleep apnea from single-lead ecg. *IEEE Transactions on Biomedical Engineering 62*, 9 (2015), 2269–2278.

[136] Vaughn, B. V., and Giallanza, P. Technical review of polysomnography. *Chest 134*, 6 (2008), 1310–1319.

[137] Vázquez et al, J.-C. Automated analysis of digital oximetry in the diagnosis of obstructive sleep apnoea. *Thorax 55*, 4 (2000), 302–307.

[138] Watanabe, E., Kiyono, K., Matsui, S., Somers, V. K., Sano, K., Hayano, J., Ichikawa, T., Kawai, M., Harada, M., and Ozaki, Y. Prognostic importance of novel oxygen desaturation metrics in patients with heart failure and central sleep apnea. *Journal of cardiac failure 23*, 2 (2017), 131–137.

[139] Whitney et al, C. W. Reliability of scoring respiratory disturbance indices and sleep staging. *Sleep 21*, 7 (1998), 749–757.

[140] Xie, B., and Minn, H. Real-time sleep apnea detection by classifier combination. *IEEE Transactions on Information Technology in Biomedicine 16*, 3 (2012), 469–477.

[141] Young, T., Peppard, P. E., and Gottlieb, D. J. Epidemiology of obstructive sleep apnea: a population health perspective. *American journal of respiratory and critical care medicine 165*, 9 (2002), 1217–1239.

[142] Zhu, K., Li, M., Akbarian, S., Hafezi, M., Yadollahi, A., and Taati, B. Vision-based heart and respiratory rate monitoring during sleep–a validation study for the population at risk of sleep apnea. *IEEE Journal of Translational Engineering in Health and Medicine 7* (2019), 1–8.

[143] Zinchuk et al, A. V. Phenotypes in obstructive sleep apnea: a definition, examples and evolution of approaches. *Sleep medicine reviews 35* (2017), 113–123.

[144] Zink, M. D., Brüser, C., Stüben, B.-O., Napp, A., Stöhr, R., Leonhardt, S., Marx, N., Mischke, K., Schulz, J. B., and Schiefer, J. Unobtrusive nocturnal heartbeat monitoring by a ballistocardiographic sensor in patients with sleep disordered breathing. *Scientific reports 7*, 1 (2017), 13175.

Curriculum vitae

Margot Deviaene was born in Waregem, Belgium on May 27, 1993. In June 2016, she obtained the Master of Science degree in Biomedical Engineering cum laude, from KU Leuven Belgium. After which, in July 2016, she joined the STADIUS center for Dynamical Signal Processing and Data Analytics in the Department of Electrical Engineering at KU Leuven as a PhD student under the supervision of Prof. dr. ir. Sabine Van Huffel, Prof. dr. Dries Testelmans and Prof. dr. ir. Carolina Varon. Her research focuses on the application of machine learning algorithms for the automatic detection and phenotyping of obstructive sleep apnea. Her work was part of the OSA+ research project, which is a collaboration between KU Leuven, UZ Leuven, TU Eindhoven, the Kempenhaeghe sleep medicine center and industrial partners. For her work on the cardiovascular phenotyping of OSA, she received the Woody Scientific sleep Award on the BASS autumn meeting in 2017 and the 3rd best poster award during the 16th Belgian National Day on Biomedical Engineering.

List of publications

Articles in internationally reviewed journals

1. **Deviaene, M.**, Borzée, P., Van Gilst, M., Van Dijk, J., Overeem, S., Buyse, B., Testelmans, D., Van Huffel, S., and Varon, C. Multilevel interval coded scoring to assess the cardiovascular status of sleep apnea patients using oxygen saturation markers. IEEE Transactions on Biomedical Engineering 67, 10 (2020), 2839-2848.

2. Lazazzera, R., **Deviaene, M.**, Varon, C., Buyse, B., Testelmans, D., Laguna, P., Gil, E., Carrault, G., Detection and Classification of Sleep Apnea and Hypopnea using PPG and SpO_2 signals. IEEE Transactions on Biomedical Engineering (2020), available in early access.

3. Varon, C., Morales, J., Lázaro, J., Orini, M., **Deviaene, M.**, Kontaxis, S., Testelmans, D., Buyse, B., Borzée, P., Sörnmo, L., et al. A comparative study of ecg-derived respiration in ambulatory monitoring using the single-lead ecg. Scientific Reports 10, 1 (2020), 1–14.

4. Milagro, J., **Deviaene, M.**, Gil, E., Lázaro, J., Buyse, B., Testelmans, D., Borzée, P., Willems, R., Van Huffel, S., Bailón, R., et al. Autonomic dysfunction increases cardiovascular risk in the presence of sleep apnea. Frontiers in Physiology 10 (2019), 620.

5. **Deviaene, M.**, Testelmans, D., Buyse, B., Borzée, P., Van Huffel, S., and Varon, C. Automatic screening of sleep apnea patients based on the SpO_2 signal. IEEE journal of biomedical and health informatics 23, 2 (2018), 607–617.

Submitted to internationally reviewed journals

1. **Deviaene, M.**, Castro, I. D., Borzée, P., Patel, A., Torfs, T., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Capacitively-coupled ECG and respiration for the detection of sleep apnea. Submitted to Physiological Measurements.

Papers in proceedings of international conferences

1. **Deviaene, M.**, Borzée, P., Buyse, B., Testelmans, D., Van Huffel, S., and Varon, C. Pulse oximetry markers for cardiovascular disease in sleep apnea. In 2019 Computing in Cardiology (CinC) (2019), IEEE, pp. Page–1.

2. Castro, I., Varon, C., Moeyersons, J., Gomez, A. V., Morales, J., **Deviaene, M.**, Torfs, T., Van Huffel, S., Puers, R., and Van Hoof, C. Data quality assessment of capacitively-coupled ecg signals. In 2019 Computing in Cardiology (CinC) (2019), IEEE, pp. Page– 1.

3. **Deviaene, M.**, Testelmans, D., Borzée, P., Buyse, B., Van Huffel, S., and Varon, C. Feature selection algorithm based on random forest applied to sleep apnea detection. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019), IEEE, pp. 2580–2583.

4. Morales Tellez, J. F., **Deviaene, M.**, Milagro, J., Testelmans, D., Buyse, B., Willems, R., Orini, M., Van Huffel, S., Bailón, R., and Varon, C. Evaluation of methods to characterize the change of the respiratory sinus arrhythmia with age in sleep apnea patients. In Proceedings of EMBC (2019), IEEE.

5. Villa, A., **Deviaene, M.**, Willems, R., Van Huffel, S., Varon, C. Are we training our heartbeat classification algorithms properly? In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019), IEEE, pp. 6363–6366.

6. Lavanga M., De Wel O., Caicedo A., **Deviaene M.**, Moeyersons J., Varon C., Bollen B., Jansen K., Ortibus E., Naulaers G., Van Huffel S. The implementation of an apnea-based perinatal stress calculator. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019), IEEE, pp. 6000-6003.

7. **Deviaene, M.**, Lázaro, J., Huysmans, D., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Sleep apnea detection using pulse

photoplethysmography. In 2018 Computing in Cardiology Conference (CinC) (2018), vol. 45, IEEE, pp. 1–4.

8. Varon C., **Deviaene M.**, Hendrikx D., Van de Putte S., Testelmans D., Buyse B., Van Huffel S. The Cardiorespiratory Graph in Sleep Apnea and Associated Comorbidities. In 2018 Computing in Cardiology Conference (CinC) (2018), vol. 45, IEEE, pp. 1–4.

9. Morales J.F., **Deviaene M.**, Milagro J., Testelmans D., Buyse B., Bailón R., Willems R., Van Huffel S., Varon C. Respiratory Sinus Arrhythmia in apnea patients with apnea associated comorbidities. In 2018 Computing in Cardiology Conference (CinC) (2018), vol. 45, IEEE, pp. 1–4.

10. **Deviaene, M.**, Varon, C., Testelmans, D., Buyse, B., and Van Huffel, S. Assessing cardiovascular comorbidities in sleep apnea patients using SpO_2. In 2017 Computing in Cardiology (CinC) (2017), IEEE, pp. 1–4.

11. Lázaro, J., Gil, E., **Deviaene, M.**, Bailón, R., Testelmans, D., Buyse, B., Varón, C., and Van Huffel, S. Pulse photoplethysmography derived respiration for obstructive sleep apnea detection. In 2017 Computing in Cardiology (CinC) (2017), IEEE, pp. 1– 4.

12. Morales, J. F., Varon, C., **Deviaene, M.**, Borzée, P., Testelmans, D., Buyse, B., and Van Huffel, S. Sleep apnea hypopnea syndrome classification in SpO_2 signals using wavelet decomposition and phase space reconstruction. In 2017 IEEE 14th international conference on wearable and implantable body sensor networks (BSN) (2017), IEEE, pp. 43–46.

Submitted book chapters

1. **Deviaene, M.**, Huysmans, D., Castro, I., Borzée, P., Testelmans, D., Buyse, B., Van Huffel, S., and Varon, C. Sleep monitoring in adults using wearables and unobtrusive technology. Submitted to Wearable/Personal monitoring devices present to future, G. Gargiulo and G. R. Naik, Eds. Springer, 2020.

Abstracts in conference proceedings

1. **Deviaene M.**, Borzee P., Buyse B., Testelmans D., Van Huffel S., Varon C., Pulse Oximetry Markers for the Assessment of Cardiovascular Disease

in Sleep Apnea, Poster presentation at the 21th annual international clinical symposium Kempenhaeghe, Kempenhaeghe, The Netherlands, Jun. 2019.

2. **Deviaene M.**, Lázaro J., Huysmans D., Testelmans D., Buyse B., Van Huffel S., Varon C., Detecting Sleep Apnea using Pulse Photoplethysmography, 7th Dutch Bio-Medical Engineering conference, Egmond aan Zee, Jan 2017.

3. **Deviaene M.**, Lázaro J., Huysmans D., Testelmans D., Buyse B., Van Huffel S., Varon C., Automatic Detection of Sleep Apnea using Pulse Photoplethysmography, 17th Belgian National Day on Biomedical Engineering, Brussels, Belgium, Nov. 2018.

4. **Deviaene M.**, Testelmans D., Buyse B., Borzée P., Van Huffel S., Varon C., Assessment of Cardiovascular Comorbidities in Obstructive Sleep Apnea Using SpO_2. Invited talk at the 20th annual international clinical symposium Kempenhaeghe, Kempenhaeghe, The Netherlands, Mar. 2018.

5. **Deviaene M.**, Testelmans D., Buyse B., Borzee P., Van Huffel S., Varon C., Detecting Cardiovascular Comorbidities In Sleep Apnea Patients Using SpO_2, In proceedings of the 16th Belgian National Day on Biomedical Engineering, Brussels, Belgium, Dec. 2017.

6. **Deviaene M.**, Testelmans D., Buyse B., Borzee P., Van Huffel S., Varon C., Assessment of Cardiovascular Comorbidities in Obstructive Sleep Apnea Using SpO_2, Proc. of the Autumn Meeting of the Belgian Association for Sleep Research and Sleep Medicine (BASS), Namur, Belgium, Nov. 2017.

7. **Deviaene M.**, Multimodal Signal Analysis for Unobtrusive Characterization of Obstructive Sleep Apnea (OSA) as part of the PhD track at the 5th Smart Systems Industry Summit (SSIS2017), Mechelen, Belgium, Oct. 2017.

8. **Deviaene M.**, Cardiac Phenotyping of OSAS patients using oxygen saturation., as part of Workshop "Towards a Less Intrusive Sleep monitoring System" at the 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN2017), Eindhoven, The Netherlands, May 2017.

9. **Deviaene M.**, De Cooman T., Lagae L., Van de Vel A., Ceulemans B., Van Huffel S. Incremental learning for online epileptic seizure detection using heart rate. In proceedings of the 15th Belgian National Day on Biomedical Engineering, Brussels, Belgium, Nov. 2016.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS
Kasteelpark Arenberg 10 box 2446
B-3001 Leuven
margot.deviaene@esat.kuleuven.be
https://www.esat.kuleuven.be/stadius/