TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Monitoring the critical newborn

Citation for published version (APA):
Varisco, G. (2023). *Monitoring the critical newborn: Towards a safe and more silent neonatal intensive care unit.* [Phd Thesis 1 (Research TU/e / Graduation TU/e), Applied Physics and Science Education]. Eindhoven University of Technology.

Document status and date:
Published: 21/03/2023

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 05. Oct. 2023

Monitoring the critical newborn

Towards a safe and more silent neonatal intensive care unit

Gabriele Varisco

Monitoring the critical newborn –

Towards a safe and more silent neonatal intensive care unit

PROEFSCRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op dinsdag 21 maart 2023 om 11:00 uur

door

Gabriele Varisco

geboren te Segrate (Milan), Italië

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Voorzitter:	prof.dr. C. Storm
Promotor:	prof.dr.ir. C. van Pul
Copromotoren:	dr.ir. X. Long
	dr. P. Andriessen (Máxima Medisch Centrum)
Promotiecommissieleden:	prof.dr.ir. M. Mischi
	prof.dr.ir. N.A.W. van Riel
	prof.dr.ir. W. Chen (Fudan University, Shanghai, China)
	prof.dr. A.H.L.C. van Kaam (Amsterdam UMC)

Het onderzoek of ontwerp dat in dit thesis wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Monitoring the critical newborn

Towards a safe and more silent neonatal intensive care unit

Gabriele Varisco

This work is part of the Alarm-Limiting AlgoRithm-based Monitoring (ALARM) project funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) with grant number 15345. The research took place in the Eindhoven MedTech Innovation Centre (e/MTIC) framework, as collaboration between the Eindhoven University of Technology, Máxima Medical Centre, and Philips Research.

The image on the cover was drawn by Samuele Gaudio

Printed by Proefschriftmaken, De Bilt, The Netherlands

A catalogue record is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-5675-5.

© Copyright 2023, Gabriele Varisco

All rights reserved. Reproduction in whole or in part is prohibited without the written consent of the copyright owner.

Summary

Premature infants are at great risk of developing severe medical complications, due to the immaturity that characterizes their physiological systems. Therefore, they are hospitalized in neonatal ICUs (NICUs), where caregivers provide them with specialized care. Physiological signals are constantly monitored to prevent deterioration and alarms are triggered in case a critical condition is found. However, different studies have shown that up to 80% of alarms sounding every day in NICUs are false or clinically irrelevant. These alarms disturb the patient and the family and can also cause alarm fatigue in caregivers, resulting in delayed or even no response to alarms. The main aim of this project is therefore to develop methods to improve patient monitoring techniques in order to reduce false alarms for a safer and more silent NICU. This project was performed within e/MTIC, a collaboration between Eindhoven University of Technology (TU/e), Máxima Medical Center and Philips and was funded by NWO-TTW in a project called ALARM.

The first part of this project focused on strategies that we implemented in the NICU of Máxima Medical Center to reduce alarms, while constantly monitoring infants' conditions to verify their safety. This part includes a study which investigated the implementation of a new alarm management solution and the optimization of alarms and protocols. This led to a significant reduction of critical alarms (30% reduction) while patients were better kept in their physiological target at the same time (35% increase in time spent in the optimal saturation target), thus improving patient safety. This study was followed by a machine learning study in which we detected when an immediate response by a caregiver was needed. To achieve this aim we used physiological signals, which included the ECG and peripheral oxygen saturation (SpO_2), as features, caregiver interaction with the patient monitors as labels for response and various combinations of training and test sets to detect caregiver response. In particular, we obtained a mean AUC of 0.93 while discriminating between stable periods and deteriorations that require an immediate response. In a third study we verified through different heart rate measures that music therapy interventions can lead to an improved autonomic regulation in premature infants and patient stability, as shown by the significant decrease in both the low-frequency and high-frequency power during the intervention compared to the preceding period.

In the second part of the project, we investigated apnea of prematurity, as this frequently occurring condition has been associated with various complications and impaired outcomes, as well as with a high number of false alarms (up to 65%). It is therefore important to keep clinically relevant apnea alarms while removing the ones that need no caregiver interaction. In our first study we optimized an algorithm for detection of

apneic events in premature infants, which led to a 10% increase in precision compared to the original algorithm with no significant change in the recall. Using the optimized algorithm an increased number of central apneas before late-onset sepsis was detected. Next, we simulated apnea of prematurity caused by immaturity of the central nervous system (i.e., central apneas) in a cardiorespiratory model in order to improve the understanding of this condition and we compared the simulations to real infant data. We found strong similarities between central apneas from patient data and simulations, which included a deeper decrease in SpO_2, with the minimum reached later in time in case of longer apneas and a faster decrease for a higher metabolic rate. In our third study we used a two-step approach for the detection of central apneas, consisting of a first detection of apneic events performed by means of the optimized algorithm for the detection of apneic events followed by the use of different machine learning models to detect whether this was a central apnea or not. The best performing model returned an AUROC of 0.88 and was found using logistic regression with elastic net penalty and leave-one-patient-out cross-validation. With this model we were able to obtain a high percentage of correct detections for all central apneas (78%) and even higher for central apneas followed by a bradycardia (94%). These are clinically the most critical types of central apneas. In the last study, a multichannel detection model based on machine learning was developed to detect central apneas directly from physiological signals. The dataset for training and testing contained not only apneic events, as in the previous study, but also of all other periods, like stable periods, periods with artefacts and periods around an apneic event from premature infants. The new model returned a precision and recall comparable to the optimized algorithm from our previous study, whilst being much faster without the two-steps, thus allowing for possible future implementation in clinical patient monitors.

Results obtained throughout this project have indicated that a safe reduction of alarms in NICUs is feasible through a workflow optimization, a revision of the protocols, the introduction of new therapy interventions as well as through the use of machine learning for the detection of the needs of caregiver response and central apnea.

Samenvatting

Te vroeg geboren kinderen lopen een groot risico op medische complicaties door de onrijpheid van hun fysiologische systemen. Daarom is het noodzakelijk om extreem of veel te vroeg geboren kinderen op te nemen op neonatale intensive care units (NICUs). Daar biedt zorgpersoneel afgestemde noodzakelijke zorg. Bovendien worden fysiologische signalen continu gemonitord om te voorkomen dat hun homeostase (volledig) ontregelt. De monitor genereert alarmen om zorgverleners te waarschuwen voor kritieke situaties.

Echter, verschillende studies hebben aangetoond dat tot 80% van de alarmen die elke dag klinken op een NICU vals zijn, of op zijn minst klinisch irrelevant. Dit soort alarmen verstoren de rust van de patiënt, de familie en ook van de zorgverleners. Ze kunnen zorgen voor desensitisatie bij zorgverleners, waardoor die mogelijk vertraagd of zelfs niet reageren op alarmen. Het voornaamste doel van dit project was daarom om methoden te ontwikkelen om de monitoring te verbeteren en daardoor valse alarmen te verminderen om een veiligere en stillere NICU op te leveren. Dit project, genaamd het ALARM project, werd uitgevoerd binnen e/MTIC, een samenwerking tussen Technische Universiteit Eindhoven (TU/e), Máxima Medisch Centrum en Philips. Het werd gefinancierd door NWO-TTW.

Het eerste deel van dit project focuste zich op strategieën die we implementeerden in de NICU van het Máxima Medisch Centrum om alarmen te verminderen, zonder de veiligheid van patiënten in het gedrang te brengen. Dit deel bevat onder andere een studie waarin we een implementatie van een nieuwe alarm management oplossing bestudeerden evenals de optimalisering van alarm protocollen. Dit leidde tot een significante vermindering van kritieke alarmen (30% reductie), terwijl patiënten voor een groter deel van de tijd in een betere gezondheidstoestand verkeerden (35% meer tijd werd doorgebracht binnen de wenselijke grenzen voor zuurstof verzadiging van het bloed). Kortom, tot een veiliger patiëntklimaat. Daarna volgde een studie gebaseerd op machinaal leren, waarin we voorspelden wanneer een acute reactie van zorgpersoneel op een alarm nodig was. Om dat doel te bewerkstelligen werden karakteristieken van de fysiologische signalen van bijvoorbeeld ECG en perifere zuurstofverzadiging (SpO_2) gebruikt als rekenmaten en de interactie van zorgverleners met de patiënt monitor werd gebruikt om het wel of niet reageren van zorgverleners op een alarm te labelen. Verschillende combinaties van training datasets en test datasets werden gebruikt om de reactie van zorgverleners te voorspellen. Specifiek vermeldenswaardig is de gemiddelde oppervlakte onder de curve van 0.93 die we verkregen toen we discrimineerden tussen stabiele perioden enerzijds en homeostatische ontregeling die een reactie van zorgverleners behoefde anderzijds.

In een derde studie verifieerden we aan de hand van maten van hartritme variabiliteit dat muziek therapie kan leiden tot verbeterde autonome regulatie en dus verbeterde patiënt stabiliteit. Dit bleek uit een significante daling in zowel laag frequente als hoog frequente spectrale componenten tijdens de muzikale interventies, in vergelijking met vóór de interventie.

In het tweede deel van dit project, bestudeerden we ademstops passend bij vroeggeboorte, ofwel apneus van prematuriteit. Dit omdat apneus van prematuriteit frequent optreden en gerelateerd worden aan verschillende andere complicaties en negatieve uitkomsten, maar óók verantwoordelijk zijn voor een groot aantal valse alarmen (tot 65%). Het is daarom belangrijk om klinisch relevante apneus goed te kunnen detecteren en die alarmen te behouden, maar om apneu alarmen die géén reactie van zorgverleners behoeven (klinisch irrelevant) te elimineren. In de eerste studie behorende bij dit deel optimaliseerden we een algoritme voor de detectie van adempauzes in te vroeg geboren kinderen. Dit resulteerde in 10% verbetering in precisie in vergelijking met het originele algoritme, zonder de herkenning van het algoritme te verminderen. Gebruik makend van dit geoptimaliseerde algoritme konden we een groter aantal centrale apneus detecteren voorafgaand aan het ontstaan van een bepaalde specifieke vorm van bloedvergiftiging die veel voorkomt bij deze populatie (late onset sepsis). Daarna simuleerden we apneus van prematuriteit met een oorsprong vanuit het centrale zenuwstelsel; centrale apneus. We simuleerden deze in een cardiorespiratoir model met als doel dit fenomeen beter te leren begrijpen en we vergeleken data uit het model met patiënten data. We vonden belangrijke vergelijkingen tussen het model en echte patiënten data. Deze waren onder andere een diepere daling van SpO_2 waarbij het diepste dal later bereikt werd in het geval van langere apneus en een snellere daling van SpO_2 bij een hogere metabole snelheid.

In de derde studie ontwikkelden we een tweeledige aanpak om de detectie van centrale apneus te verbeteren, bestaande uit eerst de detectie van adempauzes middels het geoptimaliseerde algoritme, gevolgd door het machinaal laten aanleren, middels verschillende modellen, of adempauzes ook daadwerkelijk een centrale apneu (ademstop) zijn ja of nee. Ons best presterende model, logistische regressie met elastische net penalty en 'laat één patiënt weg kruis validatie', liet een AUROC van 0.88 zien. Met dit model waren we in staat om een hoog percentage correcte identificaties van centrale apneus te laten zien (78%), vooral voor centrale apneus die op een bradycardie volgden (94%). Dit zijn klinisch gezien de meest kritische apneus. In de laatste studie ontwikkelden we een multi-kanaal detectiemodel gebaseerd op machinaal leren om centrale apneus te kunnen herkennen direct gebaseerd op het fysiologische signaal. De dataset voor het trainen en testen hiervan bevatte niet alleen adempauzes, zoals in de voorgaande studie wel het geval was, maar ook andere periodes, zoals stabiele

periodes, periodes met ruis en periodes rondom adempauzes van te vroeg geboren kinderen. Dit nieuwe model liet een precisie en herkenning zien, vergelijkbaar met het geoptimaliseerde algoritme van onze eerdere studie, maar dit model was aanzienlijk sneller, zonder dat de tweeledige aanpak nodig was. Dit maakt toekomstige toepassing in klinische patiënt monitoren mogelijk.

In dit proefschrift hebben we aangetoond dat een veilige reductie van alarmen in NICUs mogelijk wordt door optimalisatie van werkstroom, protocollen, introductie van nieuwe therapeutische interventies en door het inzetten van machinaal leren om reacties van zorgpersoneel automatisch te kunnen voorspellen en centrale apneus beter te kunnen detecteren.

Table of content

Summary v
Samenvatting vii
Table of content x
Chapter 1. Introduction 1
1.1. Introduction 3
1.2. Premature infants 3
1.3. Apnea of prematurity 6
1.4. History of neonatal intensive care units 7
1.5. Layout of neonatal intensive care units 9
1.6. Alarms and alarm fatigue 12
1.7. Aim of the thesis 13
1.8. Outline of the thesis 13
References 15
PART ONE: Alarm reduction and other optimizations of the NICU environment 21
Chapter 2. A safe reduction of alarms in NICUs is feasible 23
Abstract 24
2.1. Introduction 25
2.2. Methods 26
2.2.1.Architecture of the NICU and study population 26
2.2.2.Study phases 28
2.2.3. Workflow optimization 28
2.2.4. Alarm analysis 30
2.2.5. Survey 31
2.2.6. Statistical analysis 31
2.3. Results 32
2.3.1. Alarm analysis 32
2.3.1.1. Alarm count and duration 32
2.3.1.2. Time within SpO_2 target range 33
2.3.1.3. Nurses' response to alarms 34
2.3.1.4. Seconds of parallel alarms 34
2.3.2. Survey 34
2.4. Discussion 36
2.4.1. Alarm count and duration 36
2.4.2. Time within SpO_2 target range 37
2.4.3. Nurses' response to alarms 37

2.4.4. Seconds of parallel alarms 37
2.4.5. Survey 37
2.4.6. Additional information and limitations 38
2.5. Conclusion 39
References 40
Appendix 2.1: Nurses' responses to surveys 42
Appendix 2.2: Nurses' responses to alarms and seconds of parallel alarms 44
Chapter 3. Detecting need of caregiver response using machine learning 47
Abstract 48
3.1. Introduction 49
3.2. Related work 51
3.3. NICU monitoring workflow 52
3.4. Dataset 54
3.5. Methods 55
3.5.1. Multimodal sensors and features 55
3.5.2. Responded and deterioration events 56
3.5.3. Labeling schemes and machine learning 58
3.6. Results 60
3.6.1. Classification results 60
3.6.2. Time analysis of probabilities 64
3.7. Conclusion and future works 65
References 67
Chapter 4. Music therapy improves autonomic regulation in premature infants 71
Abstract 72
4.1. Introduction 73
4.2. Methods 74
4.2.1. Participants and communicative interventions 74
4.2.2. Data analysis 74
4.3. Results 76
4.4. Discussion 77
4.5. Conclusion 79
References 80
PART TWO: New insights into central apnea and its detection 83
Chapter 5. Facilitating annotations for central apnea in premature infants 85
Abstract 86
5.1. Introduction 87
5.2. Methods 88

5.2.1. Study design 88
5.2.2. Study datasets 89
5.2.3. Detection of central apnea-suspected events based on Lee's algorithm 90
5.2.4. Central apnea annotations by clinical experts 90
5.2.5. Data analysis and statistical analysis 92
5.3. Results 92
5.4. Discussion 95
5.5. Conclusion 96
References 97
Chapter 6. Insights into mechanisms underlying central apnea 99
Abstract 100
6.1. Introduction 101
6.2. Methods 102
6.2.1. Study population 102
6.2.2. Data analysis 102
6.2.3. Cardiorespiratory model 103
6.2.4. Model simulations 104
6.3. Results 106
6.4. Discussion 111
6.5. Conclusion 115
References 116
Appendix 6.1: Lung mechanics model 119
Appendix 6.2: Oxygen model 120
Appendix 6.3: Parameters for the lung mechanics and oxygen model 122
Chapter 7. Two-step data-driven central apnea detection using machine learning 125
Abstract 126
7.1. Introduction 127
7.2. Methods 128
7.2.1. Dataset 128
7.2.2. Annotation process 129
7.2.3. Time window for CASE detection 130
7.2.4. Signal processing 132
7.2.5. Feature extraction 132
7.2.6. Machine learning: detection models 135
7.2.7. Machine learning: validation and evaluation 136
7.2.8. Machine learning: experiments 138

7.3. Results 139
7.4. Discussion 139
7.5. Conclusion 150
References 151
Appendix 7.1: List of extracted features 154
Chapter 8. Multichannel data-driven central apnea detection using machine learning 159
Abstract 160
8.1. Introduction 161
8.2. Methods 162
8.2.1. Dataset and annotations 162
8.2.2. Signal processing and feature extraction 163
8.2.3. Window selection and undersampling 164
8.2.4. Machine learning: detection model and cross-validation 169
8.2.5. Machine learning: performance and effects evaluation 171
8.3. Results 172
8.4. Discussion 176
8.5. Conclusion 182
References 183
PART THREE: Societal impact, conclusion and future research 187
Chapter 9. Societal impact 189
9.1. Societal Impact 190
9.1.1. Impact on premature infants and their families 190
9.1.2. Impact on nurses and clinicians 191
9.1.3. Impact on industry 192
9.2. Knowledge transfer 193
References 197
Chapter 10. Conclusion and future research 199
10.1. Conclusion 201
10.2. Future research 204
References 207
Acknowledgement 208
Curriculum vitae 211
List of publications 212

CHAPTER 1

Introduction

1.1. Introduction

The birth of an infant is a moment of extreme joy and happiness for the parents and family. However, when an infant is born prematurely (i.e., before 37 complete weeks of gestational age, GA), a period of anxiety may follow due to the fragile conditions of the infant associated with prematurity [1], [2].

It has been estimated that around 15 million infants are born prematurely each year, accounting for more than one tenth of all newborn infants [3]–[5]. Premature infants can also be further classified depending on their gestational age (GA) at birth into extremely preterm (GA < 28 weeks), very preterm (GA ≥ 28 weeks and < 32 weeks) and moderate or late preterm (GA ≥ 32 week and < 37 weeks) [4]. In particular, around 16% of all premature infants are born before 32 complete weeks of GA [3]. This group of premature infants are generally characterized by more critical conditions compared to other infants due to the immaturity of their physiological systems [6]. Care for this group of premature infants is therefore typically provided in neonatal intensive care units (NICUs) where they are monitored for the full length of their stay [7]. Monitoring of vital signs, such as the ECG, respiration and peripheral oxygen saturation (SpO_2), is performed by means of patient monitoring devices which return every day a high number of alarms to alert nurses and clinicians about possible critical or alerting conditions that might occur [8], [9].

NICUs are however considered as an unnatural and stressful environment for premature infants compared to the safe environment that is their mother's womb [10]. In particular, the sound of alarms is perceived as very stressful by the infants as well as their parents [11]. In addition to this, it has been estimated that a high number of alarms sounding in NICUs consists of several false alarms as well as self-resolving ones [12], [13]. The high number of sounding alarms can result in alarm fatigue, a condition in which clinicians and nurses can become desensitized to alarm, with the risk of missing responses to a critical condition occurring in an infant [14], [15]. Methods to improve the NICU environment and to reduce the high number of alarms are therefore of paramount importance to improve the quality of care provided in NICUs. These methods should be developed considering the characteristics of the patients that were hospitalized in these hospital environments as well as the conditions and illnesses that are most frequently developed.

1.2. Premature infants

Premature infants are characterized by underdeveloped organs and very fragile conditions which can lead to several complications [5], [7]. As shown in **Figure 1.1**,

in order to prevent the occurrence of these complications constant monitoring for premature infants is needed. Complications related to premature birth are the first leading cause of neonatal death, accounting for around 35% of all deaths in this group of patients [3]. In addition to this, it has been proven that these complications can lead to further medical and neurodevelopmental consequences, such as cognitive impairment, even at a later stage in life [16]. Premature birth and its consequences cause an important psychological and financial burden for the families of these infants [17], [18]. In addition to this, it has large consequences on their future life and leads to high healthcare costs [5], [18], [19]. Infants' survival rate increases with the GA of the infants, as each week of intrauterine development increases the maturation of the different physiological systems [6]. A brief summary regarding the development at different weeks of GA for the cardiorespiratory functioning and of the autonomic nervous system (ANS), which influences the regulation of the heart rate (HR), the modulation of respiratory patterns and the gas exchange [20] is therefore presented in this paragraph.

Figure 1.1: The fragile conditions of premature infants lead to need of constant monitoring. This is possible in hospital departments known as neonatal intensive care units (NICUs). Image used under license from Shutterstock.com.

Extremely preterm infants are born at week 24 GA and are characterized by underdeveloped physiological systems [21]. These infants present an almost completely developed cardiovascular system, which is anatomically noticeable by the complete formation of the heart chambers and the main blood vessels [22]. However, an extremely immature respiratory system is found in these infants, with the terminal airways that are in an early stage of the development and lack alveolarization [5]. Production of pulmonary surfactant, a surface-active agent which will allow to reduce the surface tension in the alveoli and to prevent their collapse, starts being produced only in a very limited amount [23]. These infants require a high effort to fill their airways to provide a sufficient amount of oxygen for the full body. This complication is also known as respiratory distress syndrome (RDS) [24]. An external respiratory support as well as surfactant administration is therefore usually necessary to guarantee the survival of these infant and to prevent the occurrence of complications such as RDS. In addition to this, at this stage of life only the sympathetic nervous system (SNS) has already started its development, a process that will continue until birth [25]. Previous studies have also found a higher HR and lower heart rate variability (HRV) values, measures that provides an indication regarding the regulation of the heart and the activity of the ANS [26], [27], at birth in extremely preterm infants compared to other preterm infants born at a higher GA [2], [28].

Week 28 marks the beginning of the third trimester of pregnancy. When born at 28 weeks GA, preterm infants present more developed terminal airways with limited surfactant available, and a higher effort is requested for them to breath compared to infants born at higher GA [29]. With regards to the development of the SNS, typical structures of the adrenergic nerve terminals in the atria and ventricles can be found since this week of development [27]. The parasympathetic nervous system (PNS) has also started to develop since week 25 GA [25]. A first decrease in HR and increase in HRV measures compared to extremely preterm infants is already visible in patients born at week 28 GA, with further changes visible for infants born at higher GA [2], [28].

Moderate and late preterm are born at a minimum of 32 weeks GA. Terminal airways of these infants present a high number of alveoli, which allow for an efficient gas exchange with the blood present in the capillaries located in close proximity [5], [30]. In addition to this, at this stage of development a sufficient production of surfactant is finally reached and risks of atelectasis (i.e., partial or complete collapse of the lungs) are minimized [30]. A further decrease in HR and increase in HRV is visible through different measures for this group of patients compared to very preterm infants [2], [28], [31]. For the first time a significant increase in the parasympathetic activity can also be noticed, a result that suggests a maturation of the vagal component of the ANS.

Infants born after 37 weeks GA, often referred to as term infants, present physiological systems that are ready for birth. They are characterized by lower HR and higher HRV measures compared to premature infants [2], [27], [31]. An important increase in vagal activity can also be noticed by means of different HRV measures in term infants, an indication of the increased autonomic control that characterizes these infants.

In both term and premature infants, as shown with different HRV measures, a dominant activity of the SNS over the PNS is also noticed in the first few days after birth [27]. Social interaction and interventions performed in NICUs (e.g., skin-to-skin contact with a parent, communicative interventions, kangaroo care) as well as the time spent in different sleep states also play an important role for ANS development [25], [32], [33].

1.3. Apnea of prematurity

Apnea of prematurity is a developmental disorder occurring in premature infants, which is believed to be a manifestation of the physiologic immaturity of the respiratory control [1], [34]–[37]. Apnea of prematurity is diagnosed when frequent apneic events, characterized by cessations of breathing lasting longer than 20 s or more than 10 s in case these are followed by bradycardia or desaturation, occur in premature infants [35], [38]. Each apneic event can be classified as a central apnea (CA), caused by a cessation of respiratory drive leading to an absence of effort to breathe from the infant, an obstructive apnea, characterized by an airway obstruction and the consequent absence of respiratory flow, and a mixed apnea, which includes characteristics of both central and obstructive apneas [35], [36].

In particular, CAs account for around 20 to 40% of all the apneic events that occur during apnea of prematurity (with the remaining percentage consisting of obstructive and mixed apneas) [1], [39], [40]. As shown in **Figure 1.2**, CAs can be identified by searching for an absence of fluctuations in the respiration signal, an information that can be extracted from the ECG sensors, which are used on a daily basis in NICUs for the monitoring of premature infants. In order to detect obstructive apneas, additional sensors would be needed, such as RGB and thermal cameras, which can provide a measure of the respiratory flow of the infants for the monitoring of obstructive and mixed apneas [41].

Depending on the prevalence of a specific type of apneic event occurring in a preterm infant, different treatments can be administered. While methylxanthines (e.g., caffeine) can for instance be useful in case of CAs, as they increase minute ventilation and improve carbon dioxide sensitivity among other effects, nasal continuous positive

airway pressure is effective in case of obstructive apneas, as it contributes to opening the upper airways of the infants [35], [36].

It has also been shown that apnea of prematurity frequently occurs before the onset of very critical complications, such as sepsis, necrotizing enterocolitis, neurological impairment and brain damage [1], [42]–[45]. An accurate monitoring of apnea of prematurity is therefore relevant to assess the wellbeing of the infant, to administer an appropriate treatment depending on the different types of apneic events occurring in the infant as well as to prevent the occurrence of critical complications.

1.4. History of neonatal intensive care units

Due to the high risk of developing complications, preterm infants are hospitalized in NICUs after their birth [46]. Here their vital signs are continuously monitored by nurses and clinicians in order to stabilize their conditions and to guide the necessary treatments. It took several years to reach the standard of care that is currently present in NICUs. In addition to this, the history of NICUs is strictly connected to the one of adult intensive care units (ICUs) since each medical progress and technology adopted by one were soon investigated and applied in the other.

The first intuition for the realization of an adult ICUs date back to the 1850s, when Florence Nightingale, a British nurse and statistician, organized care for wounded soldiers in Constantinople (modern day Istanbul) during the Crimean War, as shown in **Figure 1.3**. Her work and dedication led to improved hygienic conditions in the military hospital and to a significant reduction in the death rate. In particular, her observations led to the positioning of the most critically injured soldiers closer to the nursing station so that nurses could have better monitored their conditions [32], [33].

The first key contribution to the realization of NICUs came from prof. Jean Louis Paul Denucé, who in 1857 developed the first incubator for the care of premature infants [47]. Only a few years later, in 1878, the French obstetrician Stéphane Tarnier, after observing the characteristics of chicken incubators, improved the previous design of Denucé by adding a warming chamber to warm up the infants [47], [48]. This introduction allowed to significantly decrease the death rate for very low birth weight (less than 2000 g) infants. At the end of the 18th century Martin A. Couney started to offer

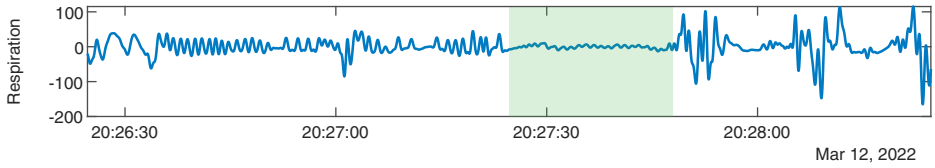


Figure 1.2: Example of central apnea (i.e., apneic event with a central origin) which is identified in green in the respiration signal of a premature infant.

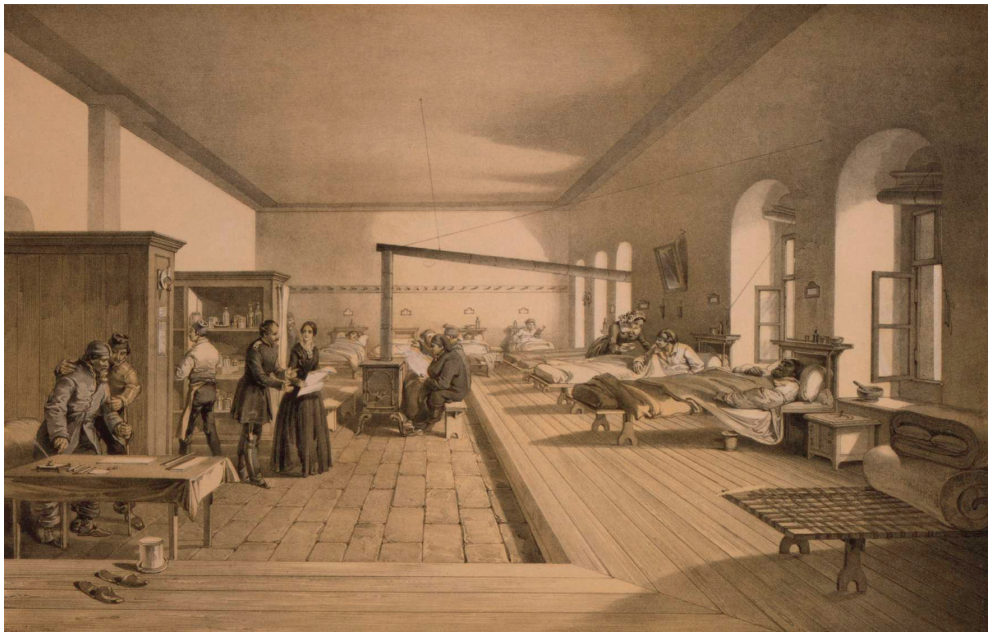


Figure 1.3: Organization of the care of the injured soldiers performed by Florence Nightingale in Constantinople during the Crimean War. Image used under license from Shutterstock.com.

specialized care for premature infants in the United States, which allowed to increase awareness for separate treatments for these patients in the whole country [47].

A rapid progress regarding the care of preterm infants is then seen since 1922, year in which the first unit dedicated exclusively to premature infant care, the Sarah Morris Premature Center, was open to the public. During 1934 the Hess oxygen box, first device able to deliver oxygen to treat respiratory distress, was developed [47], [49]. Only a few years later the Apgar Scoring System was developed to provide a standardized assessment of the infants' condition immediately after delivery [47].

In 1953, dr. Bjorn Ibsen, an anesthesiologist working in Copenhagen, came up with the idea of organizing continuous monitoring of adult patients within a dedicated ward of a hospital [50], [51]. The organization of working shifts for clinicians as well as medical students allowed to provide immediate help to the high number of patients involved in a polio epidemic, who were treated with tracheostomy followed by a treatment with positive pressure ventilation. This introduction contributed to reducing mortality from polio from over 80% to approximately 40%. Just a few years later, in 1958, dr. Max Harry Weil opened a four-bed shock ward at the Los Angeles County/University of Southern California Medical Center in the United States to improve the recognition and treatment of serious illnesses in critically ill patients [50], [52]. The success of his unit led several other hospitals in different parts of the world to replicate this way of organizing care in a dedicated ward in the immediately following years.

In the 1960s the term Newborn Intensive Care Unit was used for the first time to rename the premature nursery [47]. The 1970s and 1980s saw the introduction of new medical devices, such as neonatal ventilators, which were initially adapted from adult models and were later improved to provide a continuous positive airway pressure (CPAP) [47], [53], and new medical procedures, such as surfactant administration, which was first successfully performed to an infant by prof. Tetsuro Fujiwara [54].

Since those years, the progress achieved in the different medical and surgical fields, which also aimed at being less invasive towards the patients led to the admission of sicker patients and defined a standard for modern ICUs and NICUs [47], [50]. Modern NICUs allow to perform specialized treatments and give more attention to family centered-care by performing bonding interventions, such as kangaroo care, in which an infant is placed on a parent's chest with direct skin-to-skin contact in order to promote autonomic regulation [32], [55]. Furthermore, since NICUs are still considered as an unnatural and stressful environment for premature infants compared to the mother's womb due to the presence of lights, noise and other unnatural stimuli, researchers have focused their attention on reducing the stress experienced by the infants [10]. Recent studies have therefore for instance pointed out that communicative interventions, such as music therapy, can reduce the stress experienced by the infants and have beneficial effects on the autonomic nervous system [33], [56].

1.5. Layout of neonatal intensive care units

According to a common classification which separates ICUs based on the group of patients that can be treated, NICUs are ICUs that are dedicated to the care of infants

and preterm infants [11], [57]. Following this classification NICUs can be separated from Pediatric ICUs (PICUs), which welcome and treat children.

NICUs and nurseries can be separated based on the level of neonatal care that they can provide [58]. Being characterized by continuous availability of specialized clinicians and nurses, the availability of equipment to provide sustained life support and the possibility to provide assisted ventilation for a minimum of 24 hours, NICUs have a level 3 classification. This level is only topped by regional NICUs placed at level 4, which most often act as regional reference center to treat a specific condition of illness (e.g., retinopathy of prematurity).

Traditionally, as shown in **Figure 1.4A**, NICUs have been built with an open-bay design in order for nurses to monitor the wellbeing of all their patients at the same time. In recent years, however, the single-family room design, shown in **Figure 1.4B**, has become preferred since, being a more silent and private environment, it allows for a better brain development in the infant, better mother-child bonding while performing activities such as kangaroo care and prevents the spread of infections [11], [59], [60]. This layout also requires the presence of central nursing stations, which allows nurses to continuously keep an eye on the patients without the need to enter the patient rooms unless assistance is needed (e.g., in case of patient deterioration or feeding rounds).

In a NICU infants are usually cared for in incubators and this is particularly common for infants born very prematurely. In addition to this, NICUs are equipped with standard patient monitoring devices and devices able to support the vital conditions of their patients [11], [61]. However, due to the small size and fragile conditions of these patients, particular attention is given to the use of the least possible number of sensors, as well as to the use of less invasive solutions. Patient monitoring techniques are constantly being improved also thanks to the development of new technologies with the aim of preventing deteriorations and the use of new and less unobtrusive techniques and technologies, which include for instance radars, RGB cameras or thermal cameras to measure the respiratory flow of the infant [7], [41], [62], [63].

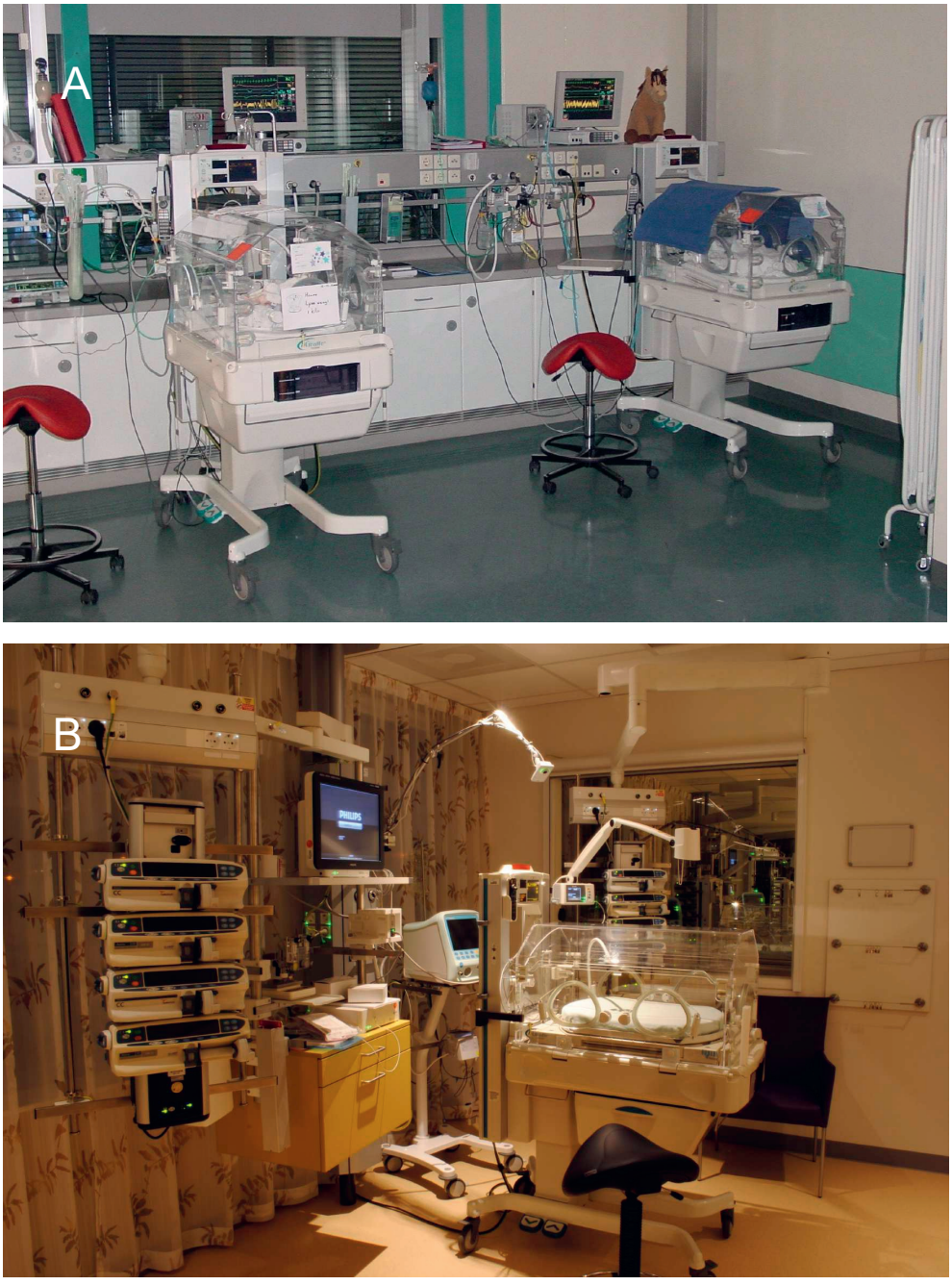


Figure 1.4: Layout for (A) an open-bay and (B) a single-family room neonatal intensive care unit (NICU). While in first case infants are all monitored together in one dedicated environment, in the latter case each room is dedicated to the monitoring of only one infant (or two/three in case twins/triplets). In addition to this, single family rooms also allow to accommodate the parents of the infant. Both images refer to the NICU of Máxima Medical Center in Veldhoven (The Netherlands), where the open-bay was present until 1999 and was later modified to a single-family room NICU.

1.6. Alarms and alarm fatigue

Medical devices generate alarms to warn clinicians and nurses that a patient might be facing a complication [64]. Typically, an alarm occurs if a certain threshold value, used to separate the optimal target range for a specific signal from a critical or alerting range, is trespassed [65], [66]. In addition to this, a delay, specifying the minimum amount of time needed by a signal to continuously exceed a threshold before an alarm is issued, and an averaging, used to apply a threshold to the average of a predefined number of samples, can also be defined as additional settings.

A high number of alarms is generated every day in the NICU environments. This number is estimated to be around 200000 per year per NICU, only considering critical alarms [61]. However, up to 80% of all the alarms and up to 65% of the alarms for apnea of prematurity are considered to be false or clinically irrelevant [12], [13], [62], [67]. False and irrelevant alarms can be caused by alarms settings, which can be not well optimized for a specific group of patients, as well as by patient movement [8], [67].

Alarms are considered a major burden in NICUs since they cause stress in both patients and parents, disturb patient sleep and can lead to increased cardiorespiratory instability [13]. In addition to this, a high alarm burden caused by false and non-actionable alarms is responsible for causing alarm fatigue in clinicians and nurses, resulting in desensitization and consequent delayed or missed responses [14], [15]. This can lead to serious consequences for the patients in case important alarms are missed and interventions are not performed. Because of this reason, alarm fatigue has been named one of the top ten health technology hazards by the ECRI institute, an organization which aims at improving the safety, quality, and cost-effectiveness of care across all healthcare settings worldwide [68]–[71].

Different studies have investigated different strategies to reduce the problem of alarm fatigue and recommended some interventions. These include improving electrode-skin contact with proper skin preparation, defining appropriate alarm settings for different patient groups or individual alarms and organizing new training session to update the hospital personnel about protocols and guidelines [72], [73]. Recent studies have also started investigating the possibility of using machine learning to provide a more accurate detection or prediction of complications. These include studies focused on an early detection of signal artifacts or specific deterioration events as well as a classification of alarms as true or false [67], [74], [75].

1.7. Aim of the thesis

The work included in this thesis had two primary objectives. The first was to develop new strategies to reduce alarm burden in NICUs, aiming at a consequent reduction of stress for the infants, the nurses and the clinicians while guaranteeing a high level of care. In order to achieve this objective, different strategies to improve alarm management in NICUs were performed and the effects of new interventions on the development of the infants were verified. The second objective consisted in the development of new data-driven strategies for a reduction of alarms associated with apnea of prematurity, with a specific focus on apneic events with a central origin, also known as central apneas (CAs) [35], [79]. Main reason for this choice is the fact that a detection of CAs, which are characterized by an absence of fluctuations in the respiratory effort, is feasible by using information extracted from the ECG sensors, which are a golden standard in the clinical practice, without the need of using any additional sensor which could require a revision of the protocols for patient monitoring in a NICU. This work was performed within the framework of the Eindhoven MedTech Innovation Center (e/MTIC), which is a collaboration between Eindhoven University of Technology (TU/e), Philips Research and Máxima Medical Center (MMC).

1.8. Outline of the thesis

This thesis is organized in different parts, following the different objectives that were previously defined. In particular, PART ONE includes the chapters related to the strategies that were developed to reduce alarm burden in NICUs. PART TWO includes instead the chapters focused on improving the understanding of central apnea and its detection. PART THREE presents the societal impact resulting from this thesis as well as the conclusion and some indication for future studies. Further indications regarding the outline of the different parts are now presented:

PART ONE: Alarm reduction and other optimizations of the NICU environment

Chapters 2-4 present the results obtained within this project with regards to the optimization of the environment of NICUs and improvement of the quality of care. In particular, **Chapter 2** presents the effects of the implementation of an alarm management system and a consequent optimization of the clinical workflow which aimed at an alarm reduction in MMC NICU. **Chapter 3** includes a study which aimed at detecting alarms that need an actual response by means of different machine learning models. **Chapter 4** presents an analysis of the effects of music therapy interventions on the heart rate variability of premature infants.

PART TWO: New insights into central apnea and its detection

Chapters 5-8 are focused on CA, a specific type of apnea of prematurity, which is associated to a high number of false alarms in the current clinical practice in NICUs. **Chapter 5** presents a study focused on the optimization of an existing algorithm for the detection of this condition and its consequent use for the detection of CAs before late-onset sepsis. **Chapter 6** includes the comparison between CAs found in a population of premature infants admitted to the NICU of MMC and simulations performed using a cardiorespiratory model. **Chapter 7** shows the results of the application of a two-step model, based on the optimized algorithm for CA detection and on machine learning, to improve detection of CAs among other apneic events. **Chapter 8** presents instead the results of the use of a one-step multichannel model based on machine learning for the detection of central apneas in a dataset which included several hours of data acquired from premature infants.

PART THREE: Societal impact, conclusion and future research

Chapter 9 includes the societal impact resulting from all the previous chapters together with a brief description of additional contributions which did not fit the main trajectory presented in this thesis. **Chapter 10** presents the conclusion for the current project as well as some indications for future research.

References

[1] J. M. Di Fiore, R. J. Martin, and E. B. Gauda, "Apnea of prematurity - Perfect storm," *Respir. Physiol. Neurobiol.*, vol. 189, no. 2, pp. 213–222, 2013.

[2] V. Golder, M. Hepponstall, S. R. Yiallourou, A. Odoi, and R. S. C. Horne, "Autonomic cardiovascular control in hypotensive critically ill preterm infants is impaired during the first days of life," *Early Hum. Dev.*, vol. 89, no. 6, pp. 419–423, 2013.

[3] H. Blencowe *et al.*, "Born Too Soon: The global epidemiology of 15 million preterm births," *Reprod. Health*, vol. 10, no. SUPPL. 1, pp. 1–14, 2013.

[4] J. A. Quinn *et al.*, "Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data," *Vaccine*, vol. 34, no. 49, pp. 6047–6056, 2016.

[5] R. E. Behrman and A. S. Butler, *Preterm birth: Causes, Consequences, and prevention*. 2007.

[6] J. S. Machado, T. S. Ferreira, R. C. G. Lima, V. C. Vieira, and D. S. de Medeiros, "Premature birth: Topics in physiology and pharmacological characteristics," *Rev. Assoc. Med. Bras.*, vol. 67, no. 1, pp. 150–155, 2021.

[7] W. Chen, S. B. Oetomo, and L. Feijs, "Neonatal monitoring: Current practice and future trends.," in *Handbook of Research on Developments in E-Health and Telemedicine*, 2010, pp. 939–961.

[8] C. McClure, S. Y. Jang, and K. Fairchild, "Alarms, oxygen saturations, and SpO2 averaging time in the NICU," *J. Neonatal. Perinatal. Med.*, vol. 9, no. 4, pp. 357–362, 2016.

[9] K. R. Johnson, J. I. Hagadorn, and D. W. Sink, "Reducing Alarm Fatigue in Two Neonatal Intensive Care Units through a Quality Improvement Collaboration," *Am. J. Perinatol.*, vol. 35, no. 13, pp. 1311–1318, 2018.

[10] D. I. Maroney, "Recognizing the potential effect of stress and trauma on premature infants in the NICU: How are outcomes affected?," *J. Perinatol.*, vol. 23, no. 8, pp. 679–683, 2003.

[11] C. Van Pul *et al.*, "Alarm management in a single-patient room intensive care units," *Recent Adv. Ambient Assist. Living – Bridg. Assist. Technol. e-Health Pers. Heal. Care*, pp. 119–132, 2015.

[12] S. Siebig, S. Kuhls, M. Imhoff, U. Gather, J. Schölmerich, and C. E. Wrede, "Intensive care unit alarms-How many do we need?," *Crit. Care Med.*, vol. 38, no. 2, pp. 451–456, 2010.

[13] C. F. Poets, "Reducing alarms in the NICU," *Arch Dis Child Fetal Neonatal*, vol. 103, no. 4, pp. F297–F298, 2018.

[14] B. D. Winters *et al.*, "Technological Distractions (Part 2): A Summary of Approaches to Manage Clinical Alarms with Intent to Reduce Alarm Fatigue," *Crit. Care Med.*, vol. 46, no. 1, pp. 130–137, 2018.

[15] R. M. Harris, J. Manavizadeh, D. J. McPherson, and L. Smith, "Do you hear bells? The increasing problem of alarm fatigue.," *Pa. Nurse*, vol. 66, no. 1, pp. 10–13, 2011.

[16] Z. Marcos, "Arriving too early," *Lancet Neurol.*, vol. 12, no. 4, pp. 332–333, 2013.

[17] C. M. Cronin, C. R. Shapiro, O. G. Casiro, and M. S. Cheang, "The impact of very low-birth-weight infants on the family is long lasting. A matched control study," *Arch Pediatr Adolesc Med*, vol. 149, no. 2, pp. 151–8, 1995.

[18] A. Lakshmanan *et al.*, "The impact of preterm birth <37 weeks on parents and families: A cross-sectional study in the 2 years after discharge from the neonatal intensive care unit," *Health Qual. Life Outcomes*, vol. 15, no. 1, pp. 1–13, 2017.

[19] J. M. Hodek, J. M. von der Schulenburg, and T. Mittendorf, "Measuring economic consequences of preterm birth - methodological recommendations for the evaluation of personal burden on children and their caregivers," *Health Econ. Rev.*, vol. 1, no. 1, pp. 1–10, 2011.

[20] T. E. Dick *et al.*, "Cardiorespiratory coupling: Common rhythms in cardiac, sympathetic, and respiratory activities," *Prog. Brain Res.*, vol. 209, pp. 191–205, 2014.
[21] R. A. Boland and J. R. Bowen, "Neonatal intensive care at 22 – 24 weeks ' gestation : How low should we go ?," vol. 27, pp. 124–126, 2018.
[22] C. M. J. Tan and A. J. Lewandowski, "The Transitional Heart: From Early Embryonic and Fetal Development to Neonatal Life," *Fetal Diagn. Ther.*, vol. 47, no. 5, pp. 373–386, 2020.
[23] M. Ball, M. Hossain, and D. Padalia, *Anatomy, Airway*. StatPearls, 2021.
[24] C. Course and M. Chakraborty, "Management of Respiratory Distress Syndrome in Preterm Infants In Wales : A Full Audit Cycle of a Quality Improvement Project," *Sci. Rep.*, pp. 1–7, 2020.
[25] S. B. Mulkey and A. J. du Plessis, "Autonomic Nervous System Development and its' Impact on Neuropsychiatric Outcome," *Pediatr Res.*, vol. 85, no. 2, pp. 120–126, 2019.
[26] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, "Heart rate variability Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, no. 17, pp. 354–381, 1996.
[27] K. Javorka *et al.*, "Heart rate variability in newborns," *Physiol. Res.*, vol. 66, no. 2, pp. S203–S214, 2017.
[28] C. J. Alonzo, V. P. Nagraj, J. V Zschaebitz, D. E. Lake, J. Randall Moorman, and M. C. Spaeder, "Heart rate ranges in premature neonates using high resolution physiologic data," *J Perinatol.*, vol. 38, no. 9, pp. 1242–1245, 2018.
[29] N. H. Hillman, S. G. Kallapur, and A. H. Jobe, "Physiology of transition from intrauterine to extrauterine life," *Clin. Perinatol.*, vol. 39, no. 4, pp. 769–783, 2012.
[30] S. Rehman and D. Bacha, *Embryology, Pulmonary*. StatPearls, 2021.
[31] E. Longin, T. Gerstner, T. Schaible, T. Lenz, and S. König, "Maturation of the autonomic nervous system : Differences in heart rate variability in premature vs . term infants Maturation of the autonomic nervous system : differences in heart rate variability in premature vs . term infants," *J. Perinat. Med.*, 2006.
[32] D. R. Kommers *et al.*, "Features of Heart Rate Variability Capture Regulatory Changes During Kangaroo Care in Preterm Infants," *J. Pediatr.*, vol. 182, pp. 92-98.e1, 2017.
[33] M. Filippa *et al.*, "Maternal Singing but Not Speech Enhances Vagal Activity in Preterm Infants during Hospitalization: Preliminary Results," *Children*, vol. 9, no. 2, pp. 1–10, 2022.
[34] J. M. Abu-Shaweesh and R. J. Martin, "Neonatal apnea: What's new?," *Pediatr. Pulmonol.*, vol. 43, no. 10, pp. 937–944, 2008.
[35] E. C. Eichenwald and Committee on Fetus and Newborn, "Apnea of Prematurity," *Pediatrics*, vol. 137, no. 1, 2016.
[36] R. E. Alvaro, "Control of breathing and apnea of prematurity," *Neoreviews*, vol. 19, no. 4, pp. e224–e234, 2018.
[37] K. Fairchild *et al.*, "Clinical associations of immature breathing in preterm infants: Part 1-central apnea," *Pediatr. Res.*, vol. 80, no. 1, pp. 21–27, 2016.
[38] J. Zhao, F. Gonzalez, and D. Mu, "Apnea of prematurity: From cause to treatment," *Eur. J. Pediatr.*, vol. 170, no. 9, pp. 1097–1105, 2011.
[39] S. Picone, R. Aufieri, and P. Paolillo, "Apnea of prematurity: challenges and solutions," *Res. Reports Neonatol.*, p. 101, 2014.
[40] N. N. Finer, K. J. Barrington, B. J. Hayes, and A. Hugh, "Obstructive, mixed, and central apnea in the neonate: Physiologic correlates," *J. Pediatr.*, vol. 121, no. 6, pp. 943–950, 1992.
[41] I. Lorato *et al.*, "Automatic separation of respiratory flow from motion in thermal videos for infant apnea detection," *Sensors*, vol. 21, no. 18, pp. 1–16, 2021.

[42] B. A. Sullivan and K. D. Fairchild, "Vital signs as physiomarkers of neonatal sepsis," *Pediatr. Res.*, vol. 91, no. 2, pp. 273–282, 2022.
[43] S. M. Gephart, J. M. McGrath, J. A. Effken, and M. D. Halpern, "Necrotizing Enterocolitis Risk," *Adv. Neonatal Care*, vol. 12, no. 2, pp. 77–87, 2012.
[44] A. Janvier, M. Khairy, A. Kokkotis, C. Cormier, D. Messmer, and K. J. Barrington, "Apnea is associated with neurodevelopmental impairment in very low birth weight infants," *J. Perinatol.*, vol. 24, no. 12, pp. 763–768, 2004.
[45] I. Zuzarte, D. Sternad, and D. Paydarfar, "Predicting apneic events in preterm infants using cardio-respiratory and movement features," *Comput. Methods Programs Biomed.*, vol. 209, p. 106321, 2021.
[46] G. Schmalisch, "Neonatal monitoring," *Capnography, Second Ed.*, no. 2012, pp. 80–95, 2011.
[47] R. C. Lussky, R. F. Cifuentes, and A. M. Siddappa, "A History of Neonatal Medicine—Past Accomplishments, Lessons Learned, and Future Challenges. Part 1—The First Century," *J. Pediatr. Pharmacol. Ther.*, vol. 10, no. 2, pp. 76–89, 2005.
[48] T. E. Cone, "Perspectives in neonatology," *Hist. Rev. Recent Adv. neonatal Perinat. Med.*, pp. 9–33, 1983.
[49] J. H. Hess, "Oxygen unit for premature and very young infants," *Am J Dis Child*, p. 47:916, 1934.
[50] J. Vincent, "Critical care – where have we been and where are we going?," *Crit. Care*, vol. 17, no. Suppl 1, pp. 1–6, 2013.
[51] D. Bennett and J. Bion, "ABC of intensive care. Organisation of fetal intensive care," *BMJ*, vol. 318, pp. 1468–70, 1999.
[52] A. Grenvik and M. R. Pinsky, "Evolution of the Intensive Care Unit as a Clinical Center and Critical Care Medicine as a Discipline," *Crit. Care Clin.*, vol. 25, no. 1, pp. 239–250, 2009.
[53] G. Gregory, J. Kitterman, R. Phibbs, W. Tooley, and W. Hamilton, "Treatment of the idiopathic respiratory distress syndrome with continuous positive airway pressure," *N Engl J Med*, no. 284, pp. 1333–40, 1971.
[54] T. Fujiwara, S. Chida, Y. Watabe, H. Maeta, T. Morita, and T. Abe, "Artificial Surfactant Therapy in Hyaline-Membrane Disease," *Lancet*, vol. 315, no. 8159, pp. 55–59, 1980.
[55] S. Kaye, "Historical trends in neonatal nursing: Developmental care and NIDCAP," *J. Perinat. Neonatal Nurs.*, vol. 30, no. 3, pp. 273–276, 2016.
[56] M. Filippa *et al.*, "Systematic review of maternal voice interventions demonstrates increased stability in preterm infants," *Acta Paediatr.*, vol. 106, no. 8, pp. 1220–1229, 2017.
[57] A. Valentin and P. Ferdinande, "Recommendations on basic requirements for intensive care units: Structural and organizational aspects," *Intensive Care Med.*, vol. 37, no. 10, pp. 1575–1587, 2011.
[58] W. D. Barfield *et al.*, "Levels of neonatal care," *Pediatrics*, vol. 130, no. 3, pp. 587–597, 2012.
[59] R. D. White, "Single-family room design in the neonatal intensive care unit-challenges and opportunities," *Newborn Infant Nurs. Rev.*, vol. 10, no. 2, pp. 83–86, 2010.
[60] R. Joshi, H. Van Straaten, H. Van De Mortel, X. Long, P. Andriessen, and C. Van Pul, "Does the architectural layout of a NICU affect alarm pressure? A comparative clinical audit of a single-family room and an open bay area NICU using a retrospective study design," *BMJ Open*, vol. 8, no. 6, pp. 1–10, 2018.
[61] C. Van Pul, H. Van De Mortel, J. Bogaart, T. Mohns, and P. Andriessen, "Safe patient monitoring is challenging but still feasible in a neonatal intensive care unit with single family rooms," *Acta Paediatr. Int. J. Paediatr.*, vol. 104, no. 6, pp. e247–e254, 2015.

[62] S. B. Amin and B. Erica, "Monitoring Apnea of Prematurity: Validity of Nursing Documentation and Bedside Cardiorespiratory Monitor," *Am J Perinatol*, vol. 30, no. 8, pp. 643–648, 2013.
[63] J. R. Williamson, D. W. Bliss, D. W. Browne, P. Indic, E. Bloch-Salisbury, and D. Paydarfar, "Using physiological signals to predict apnea in preterm infants," *Conf. Rec. - Asilomar Conf. Signals, Syst. Comput.*, pp. 1098–1102, 2011.
[64] T. A. Bach, L. M. Berglund, and E. Turk, "Managing alarm systems for quality and safety in the hospital setting," *BMJ Open Qual.*, vol. 7, no. 3, 2018.
[65] M. Görges, B. A. Markewitz, and D. R. Westenskow, "Improving alarm performance in the medical intensive care unit using delays and clinical context," *Anesth. Analg.*, vol. 108, no. 5, pp. 1546–1552, 2009.
[66] T. E. Bachman, T. E. Bachman, N. P. Iyer, C. J. L. Newth, P. A. Ross, and R. G. Khemani, "Thresholds for oximetry alarms and target range in the NICU: An observational assessment based on likely oxygen tension and maturity," *BMC Pediatr.*, vol. 20, no. 1, pp. 1–9, 2020.
[67] D. Ostojic *et al.*, "Reducing false alarm rates in neonatal intensive care: A new machine learning approach," *Adv. Exp. Med. Biol.*, vol. 1232, no. January, pp. 285–290, 2020.
[68] J. P. Keller, "Clinical alarm hazards: A top ten health technology safety concern," *J. Electrocardiol.*, vol. 45, no. 6, pp. 588–591, 2012.
[69] The Joint Commission, "Joint Commission Sentinel Event Alert: Medical Device Alarm Safety in Hospitals," *Sentin. Event Alert*, no. 50, 2013.
[70] E. Institute, "Top 10 Health Technology Hazards for 2016 - A Report from Health Devices," 2015.
[71] E. Institute, "Top 10 Health Technology Hazards for 2020 - Expert Insights from Health Devices," 2020.
[72] S. Sendelbach and M. Funk, "Alarm fatigue: A patient safety concern," *AACN Adv. Crit. Care*, vol. 24, no. 4, pp. 378–386, 2013.
[73] K. R. Johnson, J. I. Hagadorn, and D. W. Sink, "Alarm Safety and Alarm Fatigue," *Clin. Perinatol.*, vol. 44, no. 3, pp. 713–728, 2017.
[74] L. Chen *et al.*, "Using supervised machine learning to classify real alerts and artifact in online multisignal vital sign monitoring data," *Crit. Care Med.*, vol. 44, no. 7, pp. e456–e463, 2016.
[75] L. M. Eerikainen, J. Vanschoren, M. J. Rooijakkers, R. Vullings, and R. M. Aarts, "Reduction of false arrhythmia alarms using signal selection and machine learning," *Physiol. Meas.*, vol. 37, no. 8, pp. 1204–1216, 2016.
[76] H. Lee *et al.*, "A new algorithm for detecting central apnea in neonates," *Physiol. Meas.*, vol. 33, no. 1, pp. 1–17, 2012.
[77] B. D. Vergales *et al.*, "Accurate automated apnea analysis in preterm infants," *Am. J. Perinatol.*, vol. 31, no. 2, pp. 157–162, 2014.
[78] C. F. Poets, "Apnea of prematurity: What can observational studies tell us about pathophysiology?," *Sleep Med.*, vol. 11, no. 7, pp. 701–707, 2010.
[79] O. P. Mathew, "Apnea of prematurity: Pathogenesis and management strategies," *J. Perinatol.*, vol. 31, no. 5, pp. 302–310, 2011.

1

PART ONE

Alarm reduction and other optimizations of the NICU environment

CHAPTER 2

A safe reduction of alarms in NICUs is feasible

Reproduced after: Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU, Acta Paediatrica, 2021, 110 (4), 1141-1150.

Abstract

To address alarm fatigue, a new alarm management system which ensures a quicker delivery of alarms together with waveform information on nurses' handheld devices was implemented and settings optimized. The effects of this clinical implementation on alarm rates and nurses' responsiveness were measured in an 18-bed single family rooms neonatal intensive care unit (NICU).

The technical implementation of the alarm management system was followed by clinical workflow optimization. Alarms and vital parameters from October 2017 to December 2019 were analyzed. Measures included monitoring alarms, nurses' response to alarms and time spent by patients in different saturation ranges. A survey among nurses was performed to evaluate changes in alarm rate and use of protocols.

A significant reduction of monitoring alarms per patient days was detected after the optimization phase (in particular for the peripheral oxygen saturation (SpO_2) ≤ 80%, $P < 0.001$). More time was spent by infants within the optimal peripheral oxygen saturation range (88% < SpO_2 < 95%, $P < 0.001$). Results from the surveys showed that false alarms are less likely to cause an inappropriate response after the optimization phase.

The implementation of an alarm management solution and an optimization program can safely reduce the alarm burden inside of the NICU environment.

2.1. Introduction

Neonatal intensive care units (NICUs) provide high-quality care to critically ill newborn patients [1]. The majority of the patient population consists of premature infants, with immature physiological functions, at greater risk of developing serious complications, which include necrotizing enterocolitis [2], bronchopulmonary dysplasia or retinopathy of prematurity [3], [4]. They often require respiratory support from ventilators and parenteral feeding and medication via infusion pumps. To detect possibly critical events, physiological signals, which often include heart rate (HR), respiration rate, arterial blood pressure and oxygen saturation levels, are continuously monitored to detect moments in which a predefined threshold is exceeded. This leads to the generation of an alarm to alert caregivers.

Patient monitoring devices are responsible for producing a high number of alarms. One of the most common alarms in the NICU relates to peripheral oxygen saturation (SpO_2), as the use of supplemental oxygen in preterm infants is common practice and patients need to be kept in a tight range to prevent harm [5]–[8]. However, many alarms are considered false or clinically irrelevant [9], [10]. In previous studies, it was estimated that over 70% of clinical alarms do not require clinical intervention [11]. These alarms are an important concern since they both disturb patient's sleep and increase the level of stress experienced by their parents as well as the hospital personnel [12]. Alarm fatigue occurs when caregivers are exposed to large numbers of alarms, resulting in desensitization, delayed or even no response to alarms [13], [14], which can lead to serious patient harm [11], [15]. False alarms are responsible for reducing nurses' response time and trust in the alarms [16]–[18]. In particular, when performing Intensive Care Unit care in separate rooms per patient, alarm management is a safety concern, as caregivers are not continuously at the bedside of each patient. In that scenario, central monitoring and mobile alerting solutions would need to deliver alarms to the caregiver wherever they are [19].

To reduce the alarm burden and improve alarm handling in a single-family room NICU, a new alarm management solution that displays waveform information on nurses' handhelds was introduced. In a next phase, an optimization program was performed.

The aim of this study is to evaluate the effects of the implementation of an alarm management solution followed by an optimization program on the alarm rates and nurses' responsiveness to alarms over time. Since reducing alarms could result in an unsafe situation, the time spent by patients within the optimal saturation target range is also evaluated to determine the impact of this implementation on patient safety.

2.2. Methods

2.2.1. Architecture of the NICU and study population

The NICU of Máxima Medical Center (MMC, Veldhoven, The Netherlands) includes a total of nine single infant rooms, five rooms for twins and one room for a triplet, with a total of 22 beds of which 18 are typically in use, for an admission rate of approximately 380 newborn patients per year [19]. During the study period, on average 90 per cent of the patients admitted in the NICU were born prematurely with a gestational age (GA) below 32 weeks. Patient population characteristics are specified in **Table 2.1**. A waiver for this study was provided by the medical ethical committee in accordance with the Dutch law on medical research with humans (WMO). More details about the organization of the NICU and the equipment provided in each room can also be found in other works [19]–[21].

In the unit, patient monitors are connected to a network, displaying alarms both on the central monitor and other bedside monitors. When a patient is admitted to the NICU, a patient profile is chosen depending on GA as shown in **Table 2.2** which also includes the alarm limits that are currently used. Alarms in MMC NICU are classified into three different levels of priority, which include red (critical), yellow (alerting) and blue (technical) alarms [19]. Red alarms, being associated with a potentially life-threatening situation, have the highest priority and are characterized by a loud and high-frequency high repetition rate sound. Red alarms are relayed to nurses' handhelds, and in case of no response for 45 seconds, are forwarded to a pre-determined buddy nurse. Yellow alarms sound on the monitor with a softer sound and a lower repetition rate but are not communicated to the nurses' handhelds. Blue alarms are notifications of technical problems, that sound with low repetition rate, and some are relayed to the handheld in case the monitor cannot measure or detect vital signals reliably.

Nurses also have the option to silence or pause an alarm at the bedside or from the central monitor. This is usually the first action performed in MMC upon entering the room as the alarm sounds are considered a nuisance for both patients and caregivers. Silencing stops a specific alarm for 60 seconds. On the other hand, an alarm pause stops all alarms generated by a patient for 180 seconds but can be manually terminated in case a patient handling is ended earlier. Alarm pause is typically used during nurses' handling like feeding or diaper change.

Table 2.1: Implementation phases and characteristics of the populations involved in the study.

Implementation phases	Phase-0	Phase-1	Phase-2
Whole periods	October 2017 – August 2018	September 2018 – April 2019	May 2019 – December 2019
Quarters	October 2017 – December 2017	October 2018 – December 2018	October 2019 – December 2019
Alarm management solution	Philips Emergin alert management platform	Philips CareEvent	Philips CareEvent
Other implementations	-	-	Philips Mobile Caregiver app (since September 2019)
Handheld devices	Ascom I62	Ascom Myco	Ascom Myco
Servers	Central + Emergin server	CareEvent server	CareEvent server
Integration	Alarms delivered to handhelds	No configuration changes/ accept-reject alarms possible	Improved configuration, workflow and processes
Visualization from handhelds	Alarms as notifications on handhelds	Alarms and waveforms on handhelds	Optimized alarms and waveforms on handhelds
Survey for nurses' evaluation	-	First survey (November 2018)	Second survey (November 2019)

Populations (whole periods)	Phase-0	Phase-1	Phase-2
Days	336	230	225
Patients days	11.9 ± 2.3	11.0 ± 2.3	11.1 ± 2.0
Gestational age	29.0 ± 1.2	29.1 ± 1.0	29.0 ± 1.3
Post-Menstrual age	32.4 ± 1.3	31.9 ± 1.5	31.8 ± 1.3

Populations (quarters)	Phase-0	Phase-1	Phase-2
Days	92	85	90
Patients days	11.9 ± 2.3	11.0 ± 2.3	12.1 ± 1.9
Gestational age	29.1 ± 1.0	29.2 ± 1.1	28.6 ± 1.0
Postmenstrual age	32.8 ± 1.0	32.1 ± 1.4	31.4 ± 1.0

2.2.2. Study phases

Alarm handling architecture and modifications in the study period are presented in **Table 2.1**. In phase-0, monitoring alarms were sent as messages through the central monitor to an Emergin server (Philips Medical Systems), then to the central server of the messaging system (Ascom, Sweden) and from there alarms were distributed to handheld devices (Ascom i62). In phase-1, a new alarm management system, Philips CareEvent, was implemented without changing alarm settings. This solution requires just one server for alarm handling and allows for quicker delivery of alarms. It also allows the display of waveform information of all monitor signals on handhelds, for which the Ascom Myco handheld is used. In phase-2, direct access to real-time monitoring signals was implemented and alarm settings and workflow were optimized.

2.2.3. Workflow optimization

A multidisciplinary team of clinicians, nurses and engineers were engaged in an optimization program to optimize workflow and alarm settings. A literature study, an analysis of the current situation and nurse shadowing were performed to collect information about alarm pressure, nurses' behavior and policies to improve alarm handling. Next, in several work sessions, 5 areas for improvement were determined, in accordance with literature [22]: standardization of protocols, revision of alarms settings, workflow upon patient handling, training of nurses and a definition of guideline rules for alarm handling. Finally, the following changes were chosen and implemented:

1. Protocols for application, positioning and replacement of electrodes and sensors and protocol for the patients' admission and discharge from system were standardized.

2. Delay time for saturation-related alarms was increased from 10 to 20 seconds based on the findings from [7], [23], [24]. Averaging time was decreased from 10 to 4 seconds as suggested in [21], [25]. An additional red 'smart alarm' was introduced to indicate yellow alarms lasting for more than 5 minutes. All modifications to alarms are summarized in **Table 2.2**.

3. Nurses were instructed to pause alarms before starting to handle the patients to avoid generating alarms during caregiving. Additional training sessions were organized.

4. With regards to training, key users were made responsible for checking alarm management policy regularly. Nurses were trained on the new standard protocols, alarm policy and evaluation of alarm settings.

5. Nurses participated in three brainstorm sessions. The nursing group established guidelines on handling the alarms and made a '5 golden rules' poster for instruction.

This was introduced with the aim of standardizing responses to alarms and improving the workflow considering the recent changes introduced by the new alarm management solution. Explanations on how to delegate alarms to a buddy nurse and rules for choosing the right patient profile or changing it according to the needs are included in the poster. The need for the verification of the alarm limits at the beginning of each nursing rounds and response to alarms before handing the patient is addressed. Finally, guidelines on how to handle communication with the rest of the hospital personnel and patients' relatives are described.

Table 2.2: Alarm settings (thresholds and delays) assigned to patients with different gestational ages (GA) in MMC NICU.

Alarm	Parameter	GA < 26 weeks	GA 26-36 weeks	GA ≥ 37 weeks
$SpO_2 \leq 80\%$	Threshold (%)	80	80	80
	Averaging time (s)	10 → 4	10 → 4	10 → 4
	Delay (s)	10 → 20 (Total delay 20 → 24)	10 → 20 (Total delay 20 → 24)	10 → 20 (Total delay 20 → 24)
$80\% < SpO_2 \leq 88\%$	Threshold (%)	88	88	92
	Averaging time (s)	10	10	10
	Delay (s)	10	10	10
$SpO_2 \geq 95\%$	Threshold (%)	95	95	95
	Averaging time (s)	10	10	10
	Delay (s)	10	10	10
SpO_2 smart alarm	Threshold (min)	SpO_2 yellow alarms lasting ≥ 5 min		
	Averaging time (s)	No averaging		
	Delay (s)	No delay		
Bradycardia	Threshold (bpm)	80	80	60
	Averaging time (heart beats)	- Last 12 (HR ≥ 80 beats/min) - Last 4 (HR < 80 beats/min)	- Last 12 (HR ≥ 80 beats/min) - Last 4 (HR < 80 beats/min)	- Last 12 (HR ≥ 80 beats/min) - Last 4 (HR < 80 beats/min)
	Delay (s)	No delay	No delay	No delay
Heart rate (HR) low	Threshold (bpm)	100	100	100
	Averaging time (heart beats)	Last 12 (HR ≥ 80 beats/min)	Last 12 (HR ≥ 80 beats/min)	Last 12 (HR ≥ 80 beats/min)
	Delay (s)	No delay	No delay	No delay
Heart rate (HR) high	Threshold (bpm)	200	200	200
	Averaging time (heart beats)	Last 12 (HR ≥ 80 beats/min)	Last 12 (HR ≥ 80 beats/min)	Last 12 (HR ≥ 80 beats/min)
	Delay (s)	No delay	No delay	No delay

2.2.4. Alarm analysis

Alarm load was measured by the count and duration of monitoring alarms and separated for red and yellow alarms, as used in our previous study [20] and as reviewed by Johnson et al. [22]. Analyses were performed on different alarm clusters (e.g., all red monitoring alarms), to try to identify a possible general effect of changes occurred to the NICU, and on specific alarm types (e.g., $SpO_2 \leq 80\%$), to allow for a more specific interpretation of results. Alarms were counted for each day and normalized for the number of patients present in the NICU.

The time spent in target oxygen saturation range was evaluated per day by considering time spent in various ranges: below red ($SpO_2 \leq 80\%$), below target ($80\% < SpO_2 \leq 88\%$), within target ($88\% < SpO_2 < 95\%$), above target ($95\% \leq SpO_2 < 98\%$), above red ($SpO_2 \geq 98\%$).

In addition to this, we also developed new visualization tool to represent the time spent in the different target ranges for the SpO_2 signal, which we called signal target area representation. The previously discussed percentages were first computed for each day and concatenated vertically into column vectors following the order of the ranges previously indicated. When this was performed for all the days to be included in the analysis, the column vectors were concatenated horizontally in chronological order. This information was then represented using the colors red, yellow and green to indicate the different alarm ranges and the optimal target range, a choice that was performed to replicate the color codes of most of the medical devices used in the clinical practice. This visualization tool, which can be extended to other signals such as the HR, allows for instance to investigate the effect of a treatment simply by looking at how broad the colored areas are over time. It was designed to provide a direct and intuitive overview of the time spent within the different ranges and to enable the visualization across multiple days or weeks (i.e., providing an information regarding changes over time).

Analysis of nurses' responses included percentages of alarms that were silenced and the count of auto terminated and manually terminated (non-auto terminated) alarm pauses.

One important problem that characterizes each NICU is that multiple alarms can appear in the NICU at the same time, causing an increase in nurses' workload in specific daily hours. In order to provide an estimation for this problem, we introduced a new measure: seconds of parallel alarms. This was computed (for red alarms only) by considering for each hour all seconds in which a superposition between at least two alarms generated by different patients assigned to any nurse was found. All seconds of parallel alarms found within each hour were then summed and normalized considering the number of patients present in the unit. 24 values for each day (1 value per hour) included in the analysis were therefore extracted.

All the analyses were performed in Matlab (The MathWorks, Natick, United States).

2.2.5. Survey

A survey was distributed among all NICU nurses to evaluate the perception of changes occurred during phase-1 and phase-2. Part of the questions in the survey came from our previous study [26]. Surveys were made available online to all nurses on November 2018 (n = 101) and November 2019 (n = 99), a few months after the beginning of each study period. Both surveys included the same statements that they could disagree or agree with based on a five-point Likert scale. The questions aimed at identifying nurses' perceptions about alarm rate and occurrence of false alarms (28 questions), use of protocols and settings (5 questions) and training (3 questions). The survey was conducted in Dutch, native language of all nurses, and results were translated to English. No personal information was collected in the surveys. All questions delivered with survey are available in **Appendix 2.1**.

2.2.6. Statistical analysis

Three quarter periods of the same seasons (October-December, 2017-2019) were used as a representative sample of each of the three phases to compute statistical comparisons. The same quarters were chosen in order to avoid seasonal effects. Differences between phase-0 (reference) and both phase-1 and phase-2 were analyzed. Pairwise comparisons can be affected by multiplicity as the result of multiple testing [27]. However, no correction for multiplicity was included in this work since correction for type I error is not required in case multiple primary endpoints have to be verified together [28]. Natural logarithmic transformation was applied on the count of alarms and alarm pauses. A linear regression model created considering phase as the only categorical independent variable was then fitted to the log-transformed outcome. Normality assumptions of the residuals were assessed. Wilcoxon rank sum test was applied for the analysis of alarm durations to test differences between the periods.

Hotelling's T-squared test, a generalization of Student's t test, was used to compare percentages of silenced alarms and the time spent within different saturation ranges.

Seconds of parallel alarms were inflated by substantial amount of zeros since in multiple occasions no superimposed alarms were found during a daily hour. A lognormal statistical model was fitted to non-zero seconds of parallel alarms normalized using the number of patients present during that hour in the NICU. For the zero part, logistic regression was instead used to assess the proportion of zeros between different periods.

Statistical analysis was performed using R (R Core Team, Vienna, Austria).

2.3. Results

2.3.1. Alarm analysis

2.3.1.1. Alarm count and duration

An overview of the median monthly alarm count per patient per day over three years of different implementation steps is reported in **Figure 2.1**. A decrease is observed in the number of alarms in phase-2, mainly for the number of $SpO_2 \leq 80\%$ alarms, but also seasonal fluctuations, related to severeness of illness, are noticeable. The statistical analysis shown in **Table 2.3** for the quarters October to December in 2017, 2018 and 2019 showed a significant reduction in the number of all red monitoring alarms from phase-0 to phase-1 and a further decrease in phase-2. The largest decrease is seen in $SpO_2 \leq 80\%$ alarms comparing phase-0 and phase-2 (alarms per patient days 27.7 vs 13.1, $P < 0.001$). However, an increase in the number of $80\% < SpO_2 \leq 88\%$ alarms is found over time, indicating that though critical desaturation stages are not reached, lower SpO_2 values do occur. Conversely, $SpO_2 \geq 95\%$ alarms showed a significant reduction in phase-2. A small increase in the number of bradycardia and HR alarms was noticed in phase-2. However, these values remained much lower compared to those found for SpO_2 alarms. Reduced alarm duration was found for red and yellow monitoring alarms and SpO_2-related alarms, with the most significant results found in phase-2. Duration of bradycardia and HR alarms remained instead rather similar during all phases.

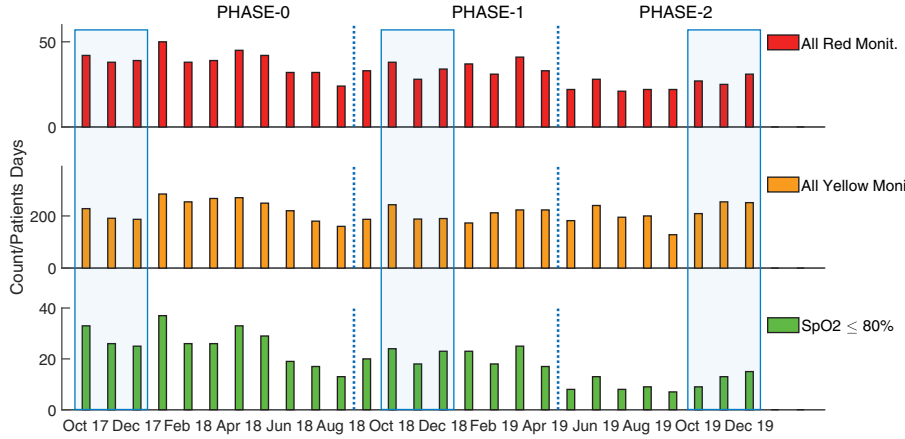


Figure 2.1: Median count of alarms/patient days for each month. Alarms for each type are counted during each day and divided by the number of patients present in the NICU on that day. Median values for each month are represented to avoid considering outliers. Dotted lines indicate the beginning and end of each phase included in this work. While this figure shows a general overview of the median values for each month, the statistical analyses for comparison between phases were performed in the same quarters highlighted in light blue (October-December, 2017-2019).

Table 2.3: Results and statistical analysis for alarm counts and durations and time spent within different SpO_2 ranges. Note: All information reported here refers to quarters (October-December) for each phase. Median values and significance are reported. Significance codes include: 0*, 0.001**, 0.01*, 0.05'. For alarm counts, the unit of measure is alarms per patient per day and the statistical method is a lognormal model. For alarm duration, the unit of measure is seconds and the statistical method is Wilcoxon rank sum. For time spent in target SpO_2 range, the unit of measure is percentages and the statistical method is Hotelling's T-squared test.**

Alarm counts	Phase-0	Phase-1	Phase-2
All red monitoring	39.5	33.4 ***	27.7 ***
All yellow monitoring	203.5	195.1	243.5 ***
$SpO_2 \leq 80\%$	27.7	20.2 ***	13.1 ***
$80\% < SpO_2 \leq 88\%$	118.9	113.7	160.0 ***
$SpO_2 \geq 95\%$	52.0	56.3	39.7 ***
Bradycardia	7.4	7.7	8.9 **
Heart rate (HR) low	14.7	14.9	19.1 ***
Heart rate (HR) high	11	12	19.2 ***

Alarm durations	Phase-0	Phase-1	Phase-2
All red monitoring	9	9	8 ***
All yellow monitoring	16	18 ***	13 ***
$SpO_2 \leq 80\%$	9	11 ***	9
$80\% < SpO_2 \leq 88\%$	18	20 ***	14 ***
$SpO_2 \geq 95\%$	23	24	19 ***
Bradycardia	5	5	5
Heart rate (HR) low	5	5	5
Heart rate (HR) high	8	9 *	9 *

Time spent in target SpO_2 range	Phase-0	Phase-1	Phase-2
Below red ($SpO_2 \leq 80$)	2.2%	1.9%	1.4% ***
Below target ($80 < SpO_2 \leq 88$)	9.3%	8.8%	7.5% ***
Within target ($88 < SpO_2 < 95$)	33.2%	35.9% *	44.9% ***
Above target ($95 \leq SpO_2 < 98$)	35.6%	35.1%	32.5% ***
Above red ($SpO_2 \geq 98$)	19.7%	18.3%	13.7% ***

2.3.1.2. Time within SpO_2 target range

Boxplots for the time spent in 5 different SpO_2 ranges are displayed in **Figure 2.2**. The percentage of time spent within SpO_2 target range ($88\% < SpO_2 < 95\%$) was found to be significantly larger in phase-2 compared to phase-0 (44.9% vs 33.2%, $P < 0.001$). Significant reduction in time spent in critical ranges outside target was found in phase-2

compared to phase-0. An important reduction was especially found for the most critical conditions, namely below red alarm limit (1.4% vs 2.2%, $P < 0.001$) and above red alarm limit (13.7% vs 19.7%, $P < 0.001$). In addition to this, the information regarding the time spent within different target ranges for the quarters October to December in 2017, 2018 and 2019 is also presented in **Figure 2.3** by means of the signal target area representation tool.

2.3.1.3. Nurses' response to alarms

The percentage of silenced alarms (shown in **Appendix 2.2**) shows no difference between the 3 phases. However, the count of alarm pauses shows an increase in phase-2. In addition, more alarms were silenced in months where alarm pressure was higher.

2.3.1.4. Seconds of parallel alarms

Periodicity can be observed in the seconds of parallel red alarms per patient per daily hour with a 2 to 3 hours period, in correspondence with the nursing rounds. Lower values for parallel alarms were noticed in both phase-1 and phase-2 (statistically significant).

2.3.2. Survey

Thirty-three nurses (32.7%) completed the survey in phase-1 and 31 nurses (31.3%) in phase-2. All questions reported in the surveys together with nurses' responses can be found in **Appendix 2.1**. Since both surveys showed high variability in responses, we chose to report only those items in which the average difference in Likert scale between pre- and post- was more than 0.5. Statistical analyses could not be performed due to the relatively low number of responses.

The main differences were perceived in the questions [3], [4], [8], [15], as shown in **Figure 2.4**. After phase-2 less nurses indicated that false alarms could reduce trust in alarms and cause an inappropriate alarm turn off. More nurses also reported that after phase-2 identification of which device is generating an alarm is more straightforward. In addition, architectural setup of the unit does not influence the way alarms are perceived and managed. Finally, in the second survey more nurses indicated that they know that in MMC NICU it is requested to document that appropriate alarms are set for each patient. For all other questions, the differences between phase-1 and 2 were smaller, though nurses reported that according to their opinion the rate of false alarms decreased over time. According to their responses, this reduction was associated with an improved interpretation of the alarm sources, an improvement in application of protocols and the use of clinical guidelines and less likelihood in missing responses to alarms. Surveys also indicated that red alarms are currently perceived as less stressful. With respect to training, the majority of nurses considered themselves well-trained in the use of handhelds. Boxplots for key results are reported in **Figure 2.4**.

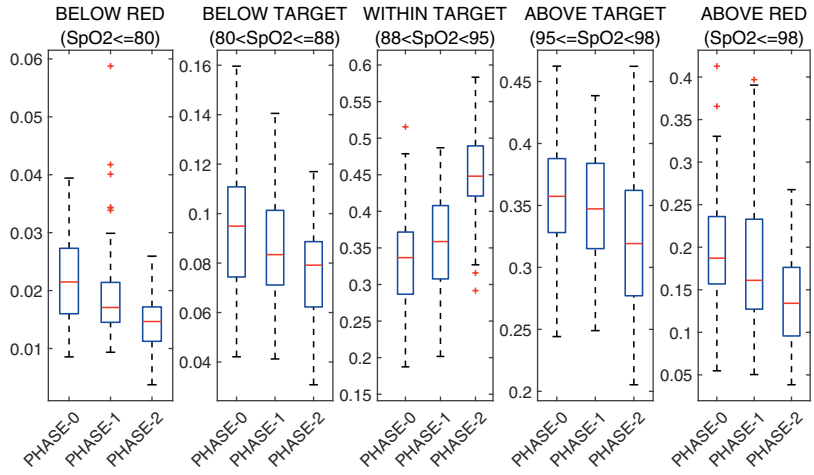


Figure 2.2: Boxplots for percentages of time spent by patients within different SpO_2 target ranges: below red ($SpO_2 \leq 80$), below target ($80 < SpO_2 \leq 88$), within target ($88 < SpO_2 < 95$), above target ($95 \leq SpO_2 < 98$), above red ($SpO_2 \geq 98$). Data reported in this figure refer to quarters that include months October-December included in each phase as for the statistical analyses.

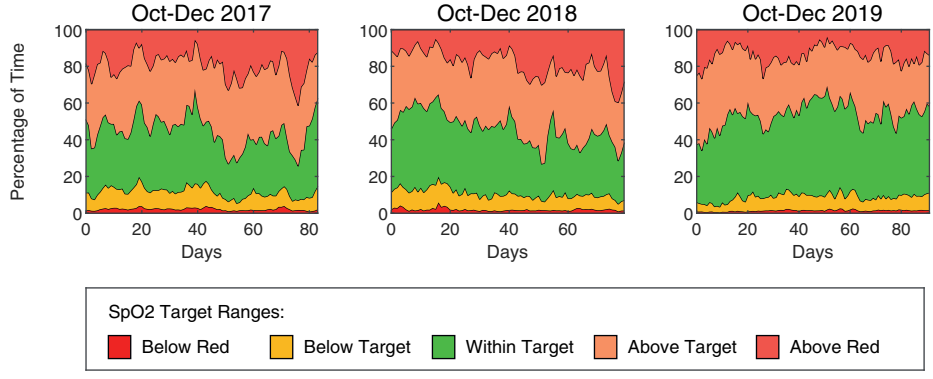


Figure 2.3: Signal target area representation for the SpO_2 signal computed by considering the following SpO_2 target ranges: below red ($SpO_2 \leq 80$), below target ($80 < SpO_2 \leq 88$), within target ($88 < SpO_2 < 95$), above target ($95 \leq SpO_2 < 98$), above red ($SpO_2 \geq 98$). Data reported in this figure refer to quarters that include months October-December included in each phase as for the statistical analyses.

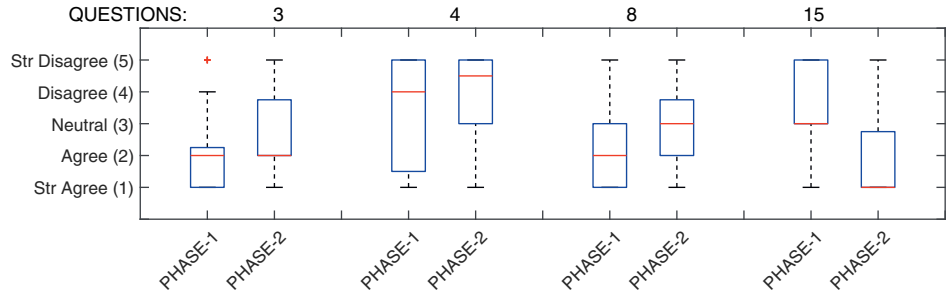


Figure 2.4: Boxplots for the surveys delivered to nurses. Nurses' responses have been analyzed and questions showing the biggest difference between the two periods have been represented. These questions are the following: (3) False alarms reduce trust in alarms and cause care givers to inappropriately turn alarms off at times other than during setup or procedures, (4) When multiple medical devices are used in a patient, it can be confusing to determine which device is in an alarm condition, (8) The architectural setup of the unit influences the way alarms are perceived and managed, (15) In my unit, there is a requirement to document that the alarms are set and are appropriate for each patient.

2.4. Discussion

This study shows that the implementation of a new alarm management solution in combination with workflow optimization can significantly reduce the number of alarms in a NICU environment, while keeping patients better into target range (saturation).

2.4.1. Alarm count and duration

$SpO_2 \leq 80\%$ alarms (red alarms) were reduced the most, particularly in phase-2, at the expense of an increase in $80\% < SpO_2 \leq 88\%$ alarms (yellow alarms). These results indicate that all changes brought to the NICU prevented patients from reaching critical state and kept them in a safer condition where the risk of developing severe complications is significantly lower. This is also supported by results for time spent within target SpO_2 range (88%-95%), which showed a significant increase in the phase-2 compared phase-0. In addition to the safer range, both patients and their parents experience less noise since yellow alarms in MMC have lower repetition times, lower frequency and a much lower volume than that of red alarms. Although a very small increase in the number of bradycardia and other HR alarms has been found in this work, we believe this would not add a lot to the alarm load. The contribution of bradycardia alarms is smaller compared to $SpO_2 \leq 80\%$ alarms (ratio 1:3, as shown in phase-0 of this work) and response to these alarms is much faster. Nurses' fast response to bradycardia alarms is indicated by the duration of HR-related alarms, which is always much shorter than for $SpO_2 \leq 80\%$ alarms.

In addition, alarm durations for both red and yellow alarms decreased significantly over time, suggesting overall faster alarm response and further noise reduction.

2.4.2. Time within SpO_2 target range

We observed a higher percentage of time spent within target range (88% < SpO_2 < 95%) in phase-1 but this percentage increases even further in phase-2, following optimization. This is also clearly visible from the signal target area representation, which clearly shows a wider area for the within target range as well as a narrower area in particular for the above red range ($SpO_2 \geq 98\%$) the following the optimization. These results indicate that both the implementation of the alarm management system and an optimization program led to increasing the time patients spent within target range, from 33% in phase 0 to 45% in phase 2, indicating improved patient safety. As the most important change was the delivery of both alarms and corresponding waveforms to nurses' handhelds, we speculate that this additional information helped nurses to adjust oxygen therapy and provide better care to patients.

2.4.3. Nurses' response to alarms

The significant increase in the use of alarm pauses in phase-2 suggests that the intervention affected nurses' habits with regard to alarm handling. We assume this an effect of training.

2.4.4. Seconds of parallel alarms

A novel measure proposed in this study is the measure of seconds of parallel alarms. Seconds of parallel alarms computed per daily hour showed that alarm pressure shows a 2-hours periodicity which corresponds to the feeding and caregiving times planned for the patients. A reduction of these high-alarm-burden periods for the different daily hours was found in phase-2, after optimization, indicating a decrease of alarm burden in MMC NICU.

2.4.5. Survey

The survey among nurses showed variability in their responses. However, as a result of the optimization, they indicated that it is less likely to inappropriately turn alarms off at times due to false alarms and misinterpret the source of alarms generated by different devices. Furthermore, the survey showed that nurses' perception on how architecture of the unit influences the perception and management of alarms has changed. This suggests that appropriate changes to protocols and settings could be helpful in improving the alarm management irrespective of the NICU architecture. These positive changes perceived by nurses can be linked to the improved clinical workflow in MMC NICU. The number of responses received from nurses for the survey is, however, limited compared to the total number of nurses in the NICU. Some reasons for the low response

rate can be found in the free participation to the survey and the fact that surveys were delivered online. Due to the low rate of responses and the possible introduction of selection bias in the analysis, these results can only be used to understand the general perception of the different changes applied to the NICU. No exact conclusions can be drawn from it since statistical analysis was not performed.

2.4.6. Additional information and limitations

This work provides a method for optimization of alarms while monitoring parameters that are indicative of patient safety. This work also investigates how the workflow optimization changes are perceived by the nurses. The content of this work can also be extended to others NICUs, irrespectively of the architecture. The most important steps in the clinical alarm management optimization require appropriate alarm settings (e.g., thresholds, delay and averaging methods) and protocols (e.g., selection of adequate patient-specific profiles depending on their age, sensor positioning and alarm escalation to buddy nurses). Nurses' training and revision of protocols is an important step to align the behavior and response of each healthcare provider. Involvement of nurses in programs as well as capturing their contribution and helping them feel engaged (e.g., creation of the '5 golden rules' poster) are important steps in ensuring clinical transformations are performed in the clinic.

Our study has several limitations. Alarms are not classified considering patients' GA and criticality. As target range depends on GA and postmenstrual age, this could affect the current general measurement of time spent in target range, in particular as infants at term-equivalent age are supposed to have a higher saturation compared to very premature infants. However, as both GA and postmenstrual age in all evaluated periods are comparable, this should not affect the observed differences between the three phases. In a previous study, nurses in our NICU indicated that they perceive bradycardia alarm to be the most reliable alarm [26]. Because of this reason no changes have been made concerning the settings of HR alarms. Nonetheless, we think that filtering out short bradycardias, as suggested in the review by Johnson et al. [22], may help to further optimize the system in the future. Also, ventilator alarms were not included in this study since the use of these devices varied over time depending on the conditions of the patients that were present in the NICU and we could stratify for these effects. Therefore, the real alarm burden is higher than reported in this study. In one of our previous studies, the alarm burden increased from 42 to 67 red alarms per patient day when alarms due to ventilation were considered [21]. Future studies could more elaborately investigate nurses' responses by analyzing escalation of alarms and the actual time needed for responding to different alarms.

This study was mainly focused on reducing red alarms therefore future works should look more into yellow alarms. Previous studies provided different ideas for reducing yellow alarms including modification of alarm thresholds, increase of alarm delays and introduction of new patient-specific profiles depending on their age [29], [30]. Other solutions would require having a better look into repeated alarm patterns.

2.5. Conclusion

The introduction of a new alarm management solution and the optimization of the clinical workflow in MMC NICU led to a reduction in the number of all red monitoring and $SpO_2 \leq 80\%$ alarms per patient day and an increase in nurses' response to alarms. Results from surveys indicated that due to the optimization, false alarms were less likely to cause an inappropriate response. This work has provided a method to decrease alarm fatigue, leading to an improved configuration of alarm workflow and processes that leads to a reduction in alarm pressure while still being safe.

References

[1] G. Schmalisch, "Neonatal monitoring," *Capnography, Second Ed.*, no. 2012, pp. 80–95, 2011.

[2] S. M. Gephart, J. M. McGrath, J. A. Effken, and M. D. Halpern, "Necrotizing Enterocolitis Risk," *Adv. Neonatal Care*, vol. 12, no. 2, pp. 77–87, 2012.

[3] J. J. Cummings and R. A. Polin, "Oxygen Targeting in Extremely Low Birth Weight Infants," *Pediatrics*, vol. 138, no. 2, pp. e20161576–e20161576, 2016.

[4] E. Khadawardi and F. Al Hazzani, "Oxygen saturation and outcomes in preterm infants the boost ii United Kingdom, Australia, and New Zealand collaborative groups," *J. Clin. Neonatol.*, vol. 2, no. 2, p. 73, 2013.

[5] J. M. Di Fiore *et al.*, "Patterns of Oxygenation, Mortality, and Growth Status in the Surfactant Positive Pressure and Oxygen Trial Cohort," *J. Pediatr.*, vol. 186, pp. 49-56.e1, 2017.

[6] W. Tarnow-Mordi and A. Kirby, "Current Recommendations and Practice of Oxygen Therapy in Preterm Infants," *Clin. Perinatol.*, no. June, pp. 1–14, 2019.

[7] B. J. Stenson *et al.*, "Oxygen saturation and outcomes in preterm infants," *N. Engl. J. Med.*, vol. 368, no. 22, pp. 2094–2104, 2013.

[8] O. D. Saugstad, "Oxygenation of the Immature Infant: A Commentary and Recommendations for Oxygen Saturation Targets and Alarm Limits," *Neonatology*, vol. 114, no. 1, pp. 69–75, 2018.

[9] S. Siebig, S. Kuhls, M. Imhoff, U. Gather, J. Schölmerich, and C. E. Wrede, "Intensive care unit alarms-How many do we need?," *Crit. Care Med.*, vol. 38, no. 2, pp. 451–456, 2010.

[10] B. J. Drew *et al.*, "Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients," *PLoS One*, vol. 9, no. 10, 2014.

[11] S. Sendelbach and M. Funk, "Alarm fatigue: A patient safety concern," *AACN Adv. Crit. Care*, vol. 24, no. 4, pp. 378–386, 2013.

[12] A. Freudenthal, M. van Stuijvenberg, and J. B. van Goudoever, "A quiet NICU for improved infants' health, development and well-being: A systems approach to reducing noise and auditory alarms," *Cogn. Technol. Work*, vol. 15, no. 3, pp. 329–345, 2013.

[13] B. D. Winters *et al.*, "Technological Distractions (Part 2): A Summary of Approaches to Manage Clinical Alarms with Intent to Reduce Alarm Fatigue," *Crit. Care Med.*, vol. 46, no. 1, pp. 130–137, 2018.

[14] R. M. Harris, J. Manavizadeh, D. J. McPherson, and L. Smith, "Do you hear bells? The increasing problem of alarm fatigue.," *Pa. Nurse*, vol. 66, no. 1, pp. 10–13, 2011.

[15] S. Casey, G. Avalos, and M. Dowling, "Critical care nurses' knowledge of alarm fatigue and practices towards alarms: A multicentre study," *Intensive Crit. Care Nurs.*, vol. 48, pp. 36–41, 2018.

[16] E. M. Petersen and C. L. Costanzo, "Assessment of Clinical Alarms Influencing Nurses' Perceptions of Alarm Fatigue," *Dimens. Crit. Care Nurs.*, vol. 36, no. 1, pp. 36–44, 2017.

[17] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomed Instrum Technol.*, pp. 46 (4), 268–277, 2012.

[18] C. P. Landrigan, "Crying wolf: False alarms and patient safety," *J. Hosp. Med.*, vol. 10, no. 6, pp. 409–410, 2015.

[19] C. Van Pul, H. Van De Mortel, J. Bogaart, T. Mohns, and P. Andriessen, "Safe patient monitoring is challenging but still feasible in a neonatal intensive care unit with single family rooms," *Acta Paediatr. Int. J. Paediatr.*, vol. 104, no. 6, pp. e247–e254, 2015.

[20] C. Van Pul *et al.*, "Alarm management in a single-patient room intensive care units," *Recent Adv. Ambient Assist. Living – Bridg. Assist. Technol. e-Health Pers. Heal. Care*, pp. 119–132, 2015.

[21] R. Joshi, H. Van Straaten, H. Van De Mortel, X. Long, P. Andriessen, and C. Van Pul, "Does the architectural layout of a NICU affect alarm pressure? A comparative clinical audit of a single-family room and an open bay area NICU using a retrospective study design," *BMJ Open*, vol. 8, no. 6, pp. 1–10, 2018.

[22] K. R. Johnson, J. I. Hagadorn, and D. W. Sink, "Alarm Safety and Alarm Fatigue," *Clin. Perinatol.*, vol. 44, no. 3, pp. 713–728, 2017.

[23] C. McClure, S. Y. Jang, and K. Fairchild, "Alarms, oxygen saturations, and SpO2 averaging time in the NICU," *J. Neonatal. Perinatal. Med.*, vol. 9, no. 4, pp. 357–362, 2016.

[24] V. V. Paine, Christine Weirich Goel, E. Ely, S. Stave, Christopher D. Stemler, M. Zander, and C. P. Bonafide, "Systematic review of physiologic monitor alarm characteristics and pragmatic interventions to reduce alarm frequency," *J Hosp Med.*, vol. 11, no. 2, pp. 136–144, 2016.

[25] M. G. and T. H. Taibo Li, Minoru Matsushima, Wendy Timpson, Susan Young, David Miedema, "Epidemiology of Patient Monitoring Alarms in the Neonatal Intensive Care Unit," *J Perinatol.*, pp. 38(8): 1030–1038, 2018.

[26] R. Joshi, H. Van De Mortel, L. Feijs, P. Andriessen, and C. Van Pul, "The heuristics of nurse responsiveness to critical patient monitor and ventilator alarms in a private room neonatal intensive care unit," *PLoS One*, vol. 12, no. 10, pp. 1–14, 2017.

[27] G. Li *et al.*, "An introduction to multiplicity issues in clinical trials: The what, why, when and how," *Int. J. Epidemiol.*, vol. 46, no. 2, pp. 746–756, 2017.

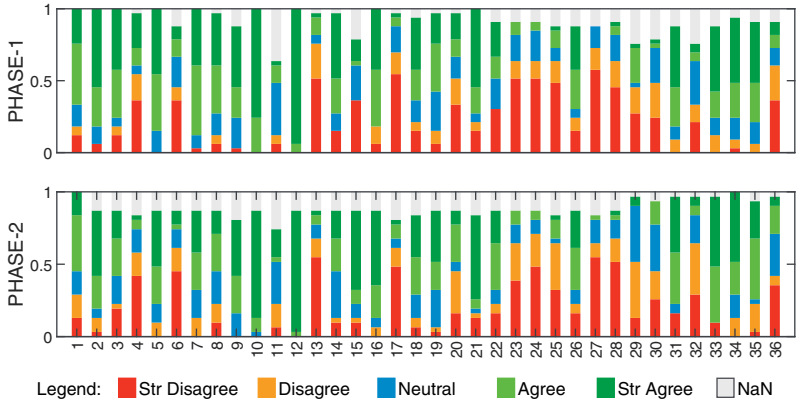
[28] European Medicines Agency, "Guideline on multiplicity issues in clinical trials," *Guidance*, vol. 44, no. December, pp. 1–15, 2016.

[29] M. Görges, B. A. Markewitz, and D. R. Westenskow, "Improving alarm performance in the medical intensive care unit using delays and clinical context," *Anesth. Analg.*, vol. 108, no. 5, pp. 1546–1552, 2009.

[30] A. K. Ketko, C. M. Martin, M. A. Nemshak, M. Niedner, and R. J. Vartanian, "Balancing the Tension Between Hyperoxia Prevention and Alarm Fatigue in the NICU," *Pediatrics*, vol. 136, no. 2, pp. e496–e504, 2015.

Appendix 2.1: Nurses' responses to surveys

Nurses' responses to surveys. Surveys were delivered on (Phase-1) and (Phase-2) and contained 36 questions organized in three topics: occurrence of false alarms (28 questions), use of protocols and settings (5 questions) and training (3 questions). Here the list of all questions organized by topic is reported:



Occurrence of false alarms: (1) False alarms often occur, (2) False alarms interrupt patient care, (3) False alarms reduce trust in alarms and cause caregivers to inappropriately turn alarms off at times other than during setup or procedures, (4) When multiple medical devices are used in a patient, it can be confusing to determine which device is in an alarm condition, (5) Smart alarms, in which different parameters, change of these parameters and signal quality are automatically assessed, would be an effective way to promote the clinical response to important alarms, (6) Employees who take over alarm management centrally and warn the responsible nurse is handy, (7) Smart alarms, in which different parameters, changes to these parameters and signal quality are automatically assessed, would be an effective way to reduce false alarms, (8) The architectural setup of the unit influences the way alarms are perceived and managed, (9) Systems that redirect alarms to handhelds are useful for promoting alarm handling, (10) Alarms with different urgency should be transmitted in different alarm sounds and / or different visual indicators, (11) The correct setting of alarm parameters and messages is too complex in existing handhelds, (12) Alarms are important to warn nurses about potential or actual changes in the condition of a patient, (13) Background noise often makes it difficult to recognize alarms, (14) Alarm sounds and/or visual indicators should be different per parameter (e.g., heart rhythm) and per device (e.g., ventilation or patient monitoring monitor), (15) In my unit, there is a requirement to document that the alarms are set and are appropriate for each patient, (16) The alarms used in my unit are adequate to inform the caregiver

about upcoming or actual deterioration of patients, (17) There have been many times when alarms could not be heard and were therefore missed, (18) Alarms from medical equipment in my unit are unique and can easily be recognized, (19) Standards, clinical guidelines and procedures for alarm management are used effectively in my unit, (20) Almost all alarms on my unit require an action / response from a healthcare provider, (21) An alarm in a life-threatening situation is clearly recognizable and leads to direct measures to get the situation under control, (22) During my service, how I experience the alarm pressure depends on the number of nurses present in the unit, (23) I often feel that I am responsible for more newborns than I can handle, (24) I often feel that I have to watch out for more newborns than I can handle, (25) When a newborn generates multiple red alarms in a short period of time (e.g., 15 minutes), I respond more slowly to the alarm, (26) If I am unable to respond to a red alarm, I feel stressed, (27) I think I often respond too slowly to a red alarm, (28) I experience red alarms as stressful.

Use of protocols and settings: (29) Too little training in alarm handhelds, (30) Difficult to understand alarm priority, (31) Disruption of the alarm handling due to non-alarm related application of the Myco handheld (such as telephone), (32) Lack of policies and guidelines on how to set alarm limits and how to adjust them specifically to the patient, (33) Frequent false alarms, these lead to reduced awareness or a reduced response to alarms if they occur.

Training: (34) I am adequately trained in the use of the monitors in my unit, (35) I am adequately trained in the use of the central station in my unit, (36) I think I need more training to be able to handle alarms adequately.

Appendix 2.2: Nurses' responses to alarms and seconds of parallel alarms

Alarm responses	Phase-0	Phase-1	Phase-2
Silenced red alarms	15.9%	19.2%	17.2%
Silenced yellow alarms	4.6%	4.5%	4.9%
Non-autoterminated pauses	1.7	1.7	4.5 ***
Autoterminated pauses	5.9	5.9	8.7 ***

Seconds of parallel red alarms	Phase-0	Phase-1	Phase-2
Non-zero data	2.7	2.8 *	2.7 ***
Zero data	6.2	9.7 ***	9.4 ***

Results and statistical analysis for nurses' responses to alarms and seconds of parallel alarms. All information reported here refers to quarters (October-December) for each phase. Median values and significance are reported. Significance codes include: 0 ***, 0.001 **, 0.01 *, 0.05 '

For silenced alarms, the unit of measure is percentages and the statistical method is Hotelling's T-squared test. For alarm pauses, the unit of measure is pauses per patients per day and the statistical method is a lognormal model. For non-zero data from seconds of parallel red alarms the unit of measure is seconds of parallel alarms per patient per hour and the statistical method is a lognormal model. For zero data from seconds of parallel red alarms, the unit of measure is hours with no parallel alarms per day and the statistical method is logistic regression.

CHAPTER 3

Detecting need of caregiver response using machine learning

Reproduced after: Listen to the Real Experts: Detecting Need of Caregiver Response in a NICU using Multimodal Monitoring Signals, ICMI '21 Companion: Companion Publication of the 2021 International Conference on Multimodal Interaction, 2021, 344–352.

Abstract

Vital signs are used in Neonatal Intensive Care Units (NICUs) to monitor the state of multiple patients at once. Alarms are triggered if a vital sign is below/above a predefined threshold. Numerous alarms sound each hour which could translate into an overload for the medical team, known as alarm fatigue. Yet many of these alarms do not require immediate clinical action of the caregivers.

In this chapter we automatically detect moments that need an immediate response (i.e., interaction with the patient) of the medical team in NICUs by using caregiver response to the patient, which is based on the interpretation of vital signs and of nonverbal cues (e.g., movements) delivered by patients. The ultimate goal of such approach is to reduce the overload of alarms while maintaining the patient safety.

We use features extracted from the ECG and pulse oximeter (peripheral oxygen saturation, SpO_2), as most unplanned interactions between patient and caregivers are due to deteriorations. Since in our unit an alarm can only be paused or silenced manually at the bedside, we used this information as a prior for caregiver response. We also propose different labeling schemes for the classification, each representative of a possible interaction scenario within the nature of our problem.

We accomplished a general detection of caregiver response with a mean AUC of 0.82. We also show that when trained only with stable and truly deteriorating windows (critical state), the classifiers can better learn the difference between alarms that need no immediate response and those that do. In addition, we present an analysis of the posterior probabilities over time for different labeling schemes and use it to speculate about the reasons behind some failure cases.

3.1. Introduction

Monitoring alarms are used in Intensive Care Units (ICU) worldwide to alert the medical team of patient deterioration [1] and should be responded to by the medical team to avoid physiological damage or even death [2]–[5]. Monitored signals generally include the heart rate (HR), blood pressure and peripheral oxygen saturation (SpO_2), among many others [2]. Critical alarms are generally accompanied by loud sounds to alert caregivers and are typically sent to a central station or sometimes to handhelds of caregivers [5], [6].

The response to critical alarms is particularly important in Neonatal Intensive Care Units (NICUs), as this fragile population has deterioration moments that tend to escalate quite rapidly. This is because preterm infants have problems regulating their physiology, as their systems are not fully developed yet and they are not as well equipped as adults to self-restore during a deterioration [6], [7].

Nonetheless, not all alarms in a (N)ICU need to be responded to immediately if they are not yet considered clinically urgent [8], [9]. Other reasons behind the lack of response for some alarms include self-patient stabilization, a large number of false or irrelevant alarms (up to 70% of total number of alarms delivered to caregivers), or work overload in the medical team [9]–[11]. Alarm fatigue is the term used for the overload of monitor alarms, resulting in desensitization to alarms in caregivers (particularly in nurses) and could potentially lead to missing important ones [2], [3], [12]. The problem complicates further in decentralized environments, such as single-family rooms [6]. Alarm fatigue can become so problematic that it has been named a medical device technology hazard [12]. On the one hand clinically relevant alarms should be handled while, to reduce alarm fatigue, alarms not clinically relevant and requiring no action should be reduced [9], [13].

In this chapter, we propose a method to detect which moments need to be responded to immediately by a caregiver, by using multimodal vital signs extracted from bedside patient monitors. Unlike current monitoring alarms which are based on values passing critical thresholds, we aim at distinguishing between important deterioration events with clinical relevance and those that can be potentially ignored. To do so, we use the interaction of the medical team with the patient monitor as silencing and pausing monitor alarms can only be done manually next to the patient. During these caregiver-patient interactions, similar to other social phenomena, the medical team learns inherently to interpret nonverbal cues from neonatal patients (e.g., movements, cry, skin tone) and reacts accordingly. We hypothesize that this additional nonverbal information from the patients, along with the other parameters extracted from vital signs, allows

the medical team to give a more accurate assessment of deterioration moments in their patients and what situations truly need a response.

This work aims therefore at improving current mechanisms for healthcare applications by revising the definition of alarms used in the NICU and substituting alarms based on thresholds with new alarms developed based on caregiver response to critical conditions recorded from their past experience. Furthermore, the solution of using multimodal vital signs together with caregiver responses as a response to the interpretation of patient signals and nonverbal cues from the patients is adopted in this work since it allowed to preserve privacy of both caregivers and patients in the NICU during the data collection and did not have an impact on working routines of caregivers. For instance, cameras were not used to record caregivers entering patient rooms as the use of these devices in intensive care setting could have resulted in a major privacy issue, a problem that was tackled with the alternative of using the interaction of the medical team with the monitor, and hence with the patient. This response can give important information on which alarms are useful and which ones can be filtered out. Reducing alarms can increase patient safety but also patient and family well-being. In addition, fewer alarms can lower stress levels for caregivers.

The main contributions of this work are:

- We automatically detect moments that need an immediate interaction of caregivers with patients in NICUs;
- We evaluate responses in moments where alarms were triggered by the bedside patient monitor, aiming to filter alarms that do not need an immediate response;
- We evaluate three different labeling schemes for training and testing of our models, based on the presence of monitor alarms and caregiver response. With each labeling scheme, we evaluate different scenarios of interaction that are inherent to the nature of our problem.

The rest of the chapter is distributed as followed: **Paragraph 3.2** summarizes the related works. **Paragraph 3.3** details some important aspects in the day-to-day of our NICU which are relevant to better understand our problem and the steps taken to solve it. The dataset used in this work is described in **Paragraph 3.4**, while **Paragraph 3.5** presents the proposed method and our different labeling schemes. **Paragraph 3.6** shows the experiments created using our labeling schemes, their results and discussion. Finally, conclusions are presented in **Paragraph 3.7**.

3.2. Related work

Some works discussed reducing alarm fatigue in NICUs by implementing delays between the moment in which a threshold is exceeded and the start of the alarm, or by using longer average periods for vital signs, to account for patient self-stabilization [14], [15]. Other heuristic approaches aimed towards wider limits for alarm thresholds [16] or customizing them to patient characteristics [13], [17]. However, these measures did not use interaction of the caregiver with the system to investigate the clinical relevance of an alarm.

Most works aiming at reducing false alarms and alarm fatigue in (N)ICUs focused on vital signs. For example, Chen et. al. [18] used supervised machine learning to detect artifacts in online vital signs from multiple sensors. Similarly, Eerikäinen et. al. [19] focused on clinical and physiological signs of arrhythmia, in order to train classifiers capable of recognizing this specific deterioration event.

Previous work studied the assessment given by the medical team, especially nurses, to critical alarms and how this influenced their reaction [8], [9], [13]. In particular, it was discussed that not all the alarms caused by vital signs exceeding thresholds triggered an immediate response by the medical team. Joshi et. al. [9] showed that nurses see all alarms as a good indicator, but they do not see the need to respond to every alarm as only a fraction is considered urgent.

More recently, Ergezen and Kol [8] provided an observational study to determine types of alarms and evaluate the response given by nurses to each type. Similarly, a rather low percentage of critical alarms (about 50%) was responded. Nevertheless, none of the above approaches used prior knowledge obtained from caregiver assessment of alarms to train detection or prediction models.

The closest work to our own was presented by Ostojic et. al. [11]. Using HR and SpO_2 with four machine learning algorithms, they classified alarms responded to by the medical team as true or false alarms. However, the staff labeled only the moments that were responded and they did so in an offline manner. In contrast, in this work we assess all alarms generated by the patient monitor, and not only those responded to. Moreover, to the best of our knowledge, we are the first to use the assessment of caregivers in the onsite response to alarms to automatically detect moments that should be responded to, aiming for alarm reduction.

3.3. NICU monitoring workflow

To better understand the nature of our problem, we describe the alarm handling workflow protocol in our NICU. There are several types of monitor alarms in a NICU, classified given their severity and nature (e.g., sensor type or technical issue). Critical monitoring alarms due to a vital sign are triggered when the signal goes below or above a threshold predefined by the medical team. The value of these thresholds varies depending on the population (i.e., very low weight, low weight) and other specific characteristics of the patient [6].

A yellow alarm is triggered when a particular vital sign of the patient is outside of the stable range, but it is not yet life-threatening. These can deteriorate further or self-stabilize. Red alarms are triggered when a vital sign goes below/above critical levels and can threaten the life of the patient. These can also self-stabilize, but this scenario is less likely. As an example for our NICU, an SpO_2 value lower than 88% triggers a yellow alarm whereas a value lower than 80% triggers a red alarm for desaturation.

All alarms produce alert sounds in the bedside monitors and are sent to a central station, for medical information and assessment. As our NICU has a single-family room arrangement, red alarms are also sent to the wearable handle of the caregiver responsible for the patient. If no response is found after 45 seconds and the red alarm does not self-stabilize, this is sent to the next available caregiver following a well-defined hierarchy. When an immediate response is needed, the caregiver will go to the patient room to provide care to the patient and handle the alarm. Responses in our NICU are only possible at the bedside, where it is possible for the caregiver to interact with the patient monitor, which then stores information about the response.

The protocols for the definition of appropriate response levels to alarms for the NICU of MMC were defined following the standards provided by the International Electrotechnical Commission in the IEC 60601-1-8, where timely responses have been deemed as necessary in order to reduce risks for the patients. Since however no further indications have been given considering the different types of alarms, medical units from hospitals all around the world have defined internal protocols that could comply with the different type of patients and illnesses that they treat. Providing a faster response to the different type of alarms in the NICU of MMC has been an important topic for the workflow optimization study presented in **Chapter 2**, where further indications about the response time for different type of alarms are provided.

The first action that is possible to perform in the patient room is to silence an alarm. Silencing means that alarms related to that sensor are silenced for 1 minute. For example,

silencing a bradycardia alarm (low heart rate) would silence all alarms related to heart rate. Another possible action is the use of a pause. When this happens, all alarms are inactive for 3 minutes. Alarms can be paused both when alarms are active and when there are no alarms. Pauses are typically used when caregivers plan to take care of a patient (e.g., diaper change, feeding or kangaroo care), knowing that these actions can disturb the sensors and lead to unnecessary alarms. Pause time can be manually stopped by caregivers before the fading time, to prevent alarms being silenced or paused when caregivers have already left the bedside. Silence and pause indicate therefore that a caregiver is close to the patient, as this action cannot be performed in our unit without being at the bedside. We call these responded events.

Figure 3.1 shows an example of alarms for HR and SpO_2 for a deteriorating patient and are subdivided into yellow and red alarms. Moments of silence/paused alarms are also shown. The interaction of the caregiver with the patient was as follows. First, an alarm indicating low SpO_2 was triggered and escalated to a desaturation (event #1 in **Figure 3.1**). This event was quite brief since soon after saturation gradually increased again. A second event (#2) came minutes later and comprehended a decrease in HR and SpO_2 followed by a desaturation. After a brief recovery, a second desaturation was noticed and silenced, allowing for the patient to recover. Finally, a third event (#3) which involved a new desaturation occurred and the caregiver paused the alarm. After the pause fading time (3 minutes) and a silenced alarm, the infant recovered and went back to stable ranges in the monitoring signals.

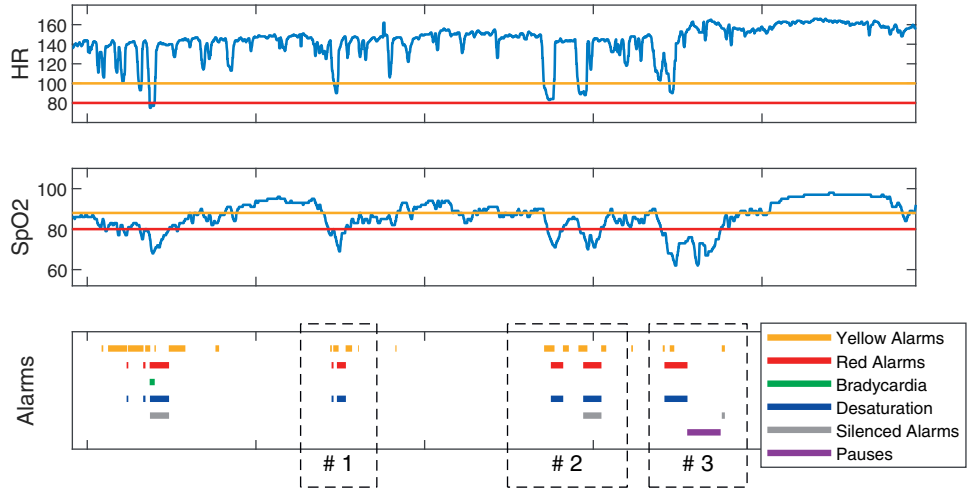


Figure 3.1: Alarm timeline for heart rate (HR), peripheral oxygen saturation (SpO_2) and caregiver interaction with the patient. In particular, three different events comprehending patient deterioration, indicated by alarms and different possible interactions, are shown.

It is also important to remind that caregiver response is caused not only by the onset of an alarm (which can be viewed at the central monitors of the NICU or through their handhelds) but also by relevant information extracted from patient history (e.g., gestational age, diseases), patient needs (e.g., diaper changes, feeding rounds) as well as the non-verbal cues (e.g., movements, cry, skin tone) delivered by the patients. This process is better illustrated in **Figure 3.2**.

3.4. Dataset

Máxima Medical Center (MMC) is a general hospital in Veldhoven (The Netherlands) with a 22-bed (18 in use), level III tertiary NICU, and an admission rate of approximately 380 newborns per year. The NICU comprises 9 single rooms, 5 twin rooms and 1 triplet room. Prematurely born infants from 24 weeks and older can be treated. All patients are continuously monitored using the Philips IntelliVue MX 800 monitor and all signals and alarms are automatically stored in a private data warehouse (PIIC iX, Data Warehouse Connect, Philips Medical Systems, Andover, MA). The sample frequency of the ECG is 250 Hz, and the HR and SpO_2 are measured every second. The patient profile that determines thresholds for the red and yellow alarms is chosen based on the gestational age of the neonate.

The data for six days for a total of 66 preterm infants was retrospectively collected from the data warehouse. It comprised the ECG, HR and SpO_2 signals, critical (red and yellow) alarms, and silenced and paused alarms. Patient demographics are presented in **Table 3.1**. The medical ethical committee of MMC provided a waiver for this study, in accordance with the Dutch law on medical research with humans (WMO).

About 5% of the time consisted of yellow alarms, while 0.8% were red alarms. In addition, 1.6% of the time were either paused or silenced by the medical team. It was thus clear that we were dealing with a problem of imbalanced classes. As such, the training of the classifiers and the metrics for its evaluation had to be defined accordingly.

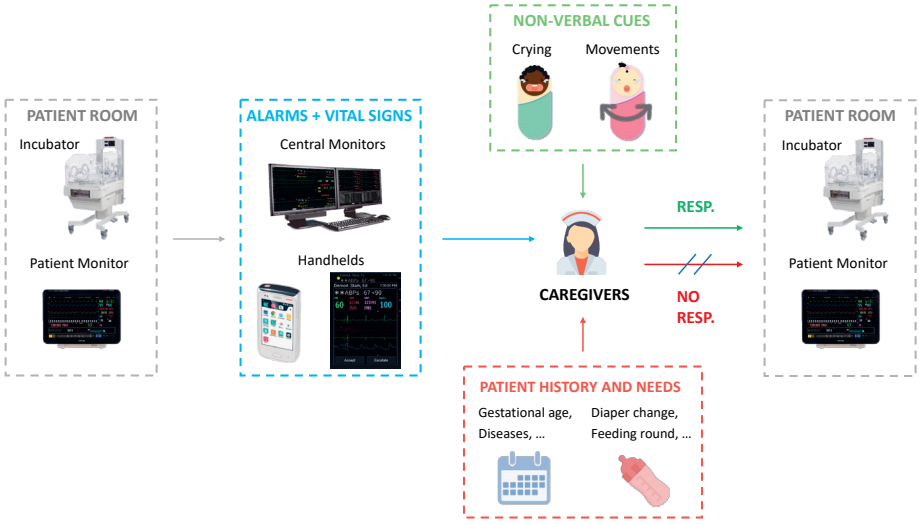


Figure 3.2: Information leading to caregiver response or no response include alarms (sounding at the central monitors as well as on the handhelds carried by caregivers), patient history (e.g., gestational age, diseases), patient needs (e.g., diaper changes, feeding rounds) and non-verbal cues (e.g., cry, movements, skin tone) delivered by the patients. In case a response is needed, caregivers are instructed to interact with the patient monitor to stop alarms from sounding and to prevent new ones from occurring before handling the patient.

3.5. Methods

Here we describe our approach to detect the need for caregiver response, based on clinically readily available data in the patient monitor. First, we describe the used sensors and the multimodal features derived from them. Then, we provide a definition for responded events and deterioration events and describe how they differ from raw critical monitoring alarms. Finally, we present our different labeling schemes and the classifiers used in our experiments.

3.5.1. Multimodal sensors and features

We used two sensors in this work: 1) the ECG, to measure the electrical activity of the heart, and 2) the pulse oximeter for the SpO_2. These sensors were chosen because the majority of critical alarms in a NICU can be extracted from them [17]. A critical alarm is defined as an alarm that requires immediate action and for premature infants, a low heart rate (<80 bpm) and a low saturation (<80%) are considered critical alarms. We derived three different feature groups: 1) HR-based, 2) heart-rate-variability (HRV)-based and 3) SpO_2-based (see **Table 3.2**).

An RR-interval is the time elapsed between two successive peaks in the ECG, called R-peaks. The HR was determined from the RR-interval as the number of beats (peaks) per minute. This signal is pre-extracted in most modern patient monitors and it was the case also for the one included in our dataset. The Philips IntelliVue MX 800 monitor adds a series of noise filtering and signal enhancing proprietary methods for a better estimation of the HR. We used the mean and the variance of the HR as our features for this modality since the variance in the HR is indicative of incoming deterioration [20]. HRV features were derived from this RR-interval signal [21], [22]. In this work, we computed the mean RR-interval value within a window (mean-RR), the standard deviation of the RR-intervals within a window (SDNN) and the root mean square of successive differences between adjacent RR-intervals (RMSSD) [22]. Finally, we calculated the mean and variance of the SpO_2.

Table 3.1: Patient demographics for our study.

Characteristic	Median	25^{th} perc.	75^{th} perc.
Gestational age (weeks)	29.4	28.7	30.7
Birth weight (g)	1260.0	1065.0	1417.5

Table 3.2: Used features per signal type. HRV: heart rate variability. SpO_2: peripheral oxygen saturation. SDNN: Standard deviation of RR-intervals. RMSSD: Root mean square of successive differences between adjacent RR-intervals.

Sensor	Feature Group	Features
ECG	Heart rate	Mean, variance
	Heart rate variability (HRV)	Mean-RR, SDNN, RMSSD
Pulse oximeter	SpO_2	Mean, variance

These statistics were used since they have proven to be informative of deteriorations and correlated to the infants' development, allowing for an easier transfer of knowledge about results to caregivers [23]–[25]. Since previous studies indicated that nurses prioritize responses to more vulnerable infants, these features might therefore be representative of the key information used by nurses to decide whether or not and how fast to provide a response [9]. A summary of all the features used in this study is presented in **Table 3.2**.

3.5.2. Responded and deterioration events

In our work pipeline, we first found the alarms. Next, we clustered the silenced and paused alarms that were within 3 minutes or less from each other. This was performed since most interactions between the caregivers and the patient constitute a long sequence

of alarms without separations. Furthermore, 3 minutes is also the longer fading time for pause/silencing in our system. Once the silenced and paused alarms were clustered into a single event, we investigated whether critical red and yellow alarms were triggered within the event. Critical alarms that were closely placed together were clustered in the same way described for silenced/paused alarms.

We then defined as a responded event any cluster of closely placed silenced and/or paused alarms and this was set as our prior for caregiver response. A deterioration event was instead defined as a responded event overlapping with or in close proximity to a red or/and yellow alarm. This solution allowed to analyze the characteristics of such events (e.g., a desaturation only, a desaturation with bradycardia, an arrhythmia). We added additional 30 seconds at the beginning and end of each responded event, to account for the time needed for a caregiver the reach the bedside in the patient room and provide a response and for stabilization after the event, respectively. In addition to this, since a pause inactivates all alarms, we had to assess the seconds before a pause.

The top section of **Figure 3.3** presents the streams of raw red and yellow alarms (indicated in yellow), alarms that were paused or silenced (represented in red) and responded events that resulted after the clustering scheme (in blue) as an example. It is clear from this figure that while critical alarms were triggered, these were not always responded.

A sliding window approach was then used to perform machine learning analyses, as represented in the bottom section of **Figure 3.3**. A window size of T = 1 min and no overlaps were selected to extract all the previously described features. This choice allowed us to consistently deal with windows of the same length and long enough to include alarms and caregiver response (since responses, in case these are present, are usually happening in less than 30 sec) while avoiding the inclusion of different consecutive alarms.

Windows including at least one critical alarm within them were classified as having an alarm (A = 1). Silenced/paused windows were clustered together to create the final responded events and to define windows with responses (R = 1). The stream of alarms (represented in red in **Figure 3.3**) and responses (represented in blue) led therefore to four possible labels per each window of length T:

- No alarm, no response (AR = 00) (i.e., stable patient);
- Alarm present, no response (AR = 10) (i.e., alarm with no response);
- No alarm, response present (AR = 01) (i.e., responded event);
- Alarm present, response present (AR = 11) (i.e., deterioration event).

3.5.3. Labeling schemes and machine learning

Due to the presence of a label for both alarms and responses in each window we decided to use three different labelling schemes in this work: complete, strict and AlarmsOnly.

Complete labelling was based only on the information regarding caregiver response, irrespectively of the presence of alarms (i.e., R = 0 vs. R = 1). With this labelling scheme we aimed at predicting windows that needed a response.

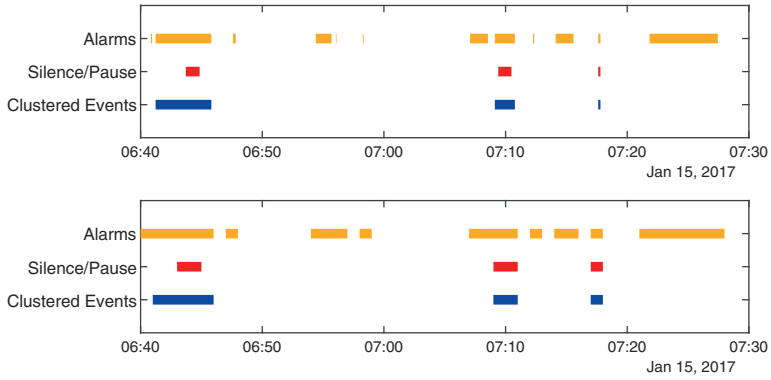


Figure 3.3: Clustering of silenced/paused alarms useful for the creation of responded events considering raw silenced/paused alarms (top) and 1 min long window (bottom). Deterioration events were defined as responded events that also had at least an alarm (red or yellow) within them.

In the strict labelling only windows in which the patient was stable (i.e., AR = 00) and deterioration events (i.e., AR = 11) were considered. All other windows were discarded. This labelling scheme aimed at verifying whether a classifier can learn the difference between a stable and a deteriorating infant, which can be translated to the identification of moment in which caregivers should provide help to patients during a deterioration event.

Finally, in the AlarmsOnly labelling scheme only windows including alarms (i.e., AR = 10 vs. AR = 11) were considered, in order to verify whether a classifier can identify moments in which a response is necessary. This was designed with the specific aim of reducing false alarms, thus reducing alarm fatigue.

Using these three labeling schemes we created the six different experiments summarized in **Table 3.3**. Experiment 1 was selected with the aim of detecting caregiver response (i.e., detection of responded events), irrespectively of the presence of alarms. In this experiment

the complete labelling was used for both the training and test set. Experiment 2 was used to see if classifiers can learn the difference between a stable patient (i.e., AR = 00) and one with a clinically relevant deterioration (i.e., AR = 11), which led to the choice for a strict labelling for both the training and test set. Experiment 3 was set to see if it is possible to discriminate between windows that need a response among all those that included an alarm triggered by the patient monitor (i.e., AR = 10 vs. AR = 11). For this experiment the AlarmsOnly labelling was used for both the training and test set.

Experiments 4 to 6 made use of different combination of labeling schemes for training and testing. With experiment 4 the impact of learning the true separation between a deterioration and a stable patient on the detection of all responded events (i.e., training with strict, testing with complete) was evaluated. With experiment 5, we assessed whether this same difference influenced the detection of windows having an alarm triggered by the patient monitor that needed a response (i.e., training with strict, testing with AlarmOnly). Finally, with experiment 6 we evaluated whether training using responses only might be beneficial for the detection of windows having an alarm that needed a response (i.e., training with complete, testing with AlarmOnly).

It is also important to underline that while the main goal of this work is to detect windows that should be responded (i.e., responded events), from a clinical point of view the most relevant experiments are those for which the AlarmOnly label was used during testing (3, 5 and 6) since understanding the difference between alarms that needed a response and those that didn't require one could lead to the reduction of false alarms and alarm fatigue.

For each experiment, we applied a leave-5-subjects-out cross-validation using the same training and test data in all experiments for a fair comparison. All windows for 5 subjects (selected at random without replacement) were therefore left out for testing in each fold, while the rest of windows were used for training. No separate validation set was defined to perform the experiments included in this chapter.

Table 3.3: Performed experiments. Each experiment represents a relevant combination of the different labelling schemes for training and testing.

Experiment	Used labeling			
	Training	Windows labels	Testing	Windows labels
1	Complete	R = 0 vs. R = 1	Complete	R = 0 vs. R = 1
2	Strict	AR = 00 vs. AR = 11	Strict	AR = 00 vs. AR = 11
3	AlarmsOnly	AR = 10 vs. AR = 11	AlarmsOnly	AR = 10 vs. AR = 11
4	Strict	AR = 00 vs. AR = 11	Complete	R = 0 vs. R = 1
5	Strict	AR = 00 vs. AR = 11	AlarmsOnly	AR = 10 vs. AR = 11
6	Complete	R = 0 vs. R = 1	AlarmsOnly	AR = 10 vs. AR = 11

We used a lasso logistic regressor (logistic regressor with L1 penalty) and a nearest mean classifier (also known as nearest centroid classifier) for all our experiments. We selected these to assess the problem with a linear separation and a clustering-based approach, respectively. Another reason for the choice of lasso logistic regression and nearest mean lay in the simplicity that characterizes these two classifiers compared to other methods (e.g., neural networks). These classifiers allow in fact to explain the importance of each feature, which is critical for the transfer of knowledge about the results achieved with this work to caregivers. To consider the imbalance in our dataset during training, we weighted the contribution of each type of window during the optimization of the objective function [26]. For the evaluation, we used the Receiver Operating Characteristic curve (ROC) and the area under them (AUROC value). Unlike accuracy, these metrics are more suited to handle imbalanced data [26], [27]. In addition, we present confusion matrices for further evaluation.

3.6. Results

3.6.1. Classification results

Mean (± standard deviation per fold) AUROC values for all the experiments are summarized in **Table 3.4** while the mean ROC curves and both classifiers are shown in **Figure 3.4**.

Table 3.4: Mean AUROC values (± standard deviation between folds) of the detection of caregiver response for the different cases.

Experiment	Training	Testing	AUROC logistic regr.	AUROC nearest mean
1	Complete	Complete	0.82 ± 0.03	0.81 ± 0.03
2	Strict	Strict	0.93 ± 0.04	0.94 ± 0.03
3	AlarmsOnly	AlarmsOnly	0.78 ± 0.07	0.77 ± 0.07
4	Strict	Complete	0.79 ± 0.04	0.80 ± 0.04
5	Strict	AlarmsOnly	0.77 ± 0.07	0.78 ± 0.07
6	Complete	AlarmsOnly	0.78 ± 0.07	0.78 ± 0.07

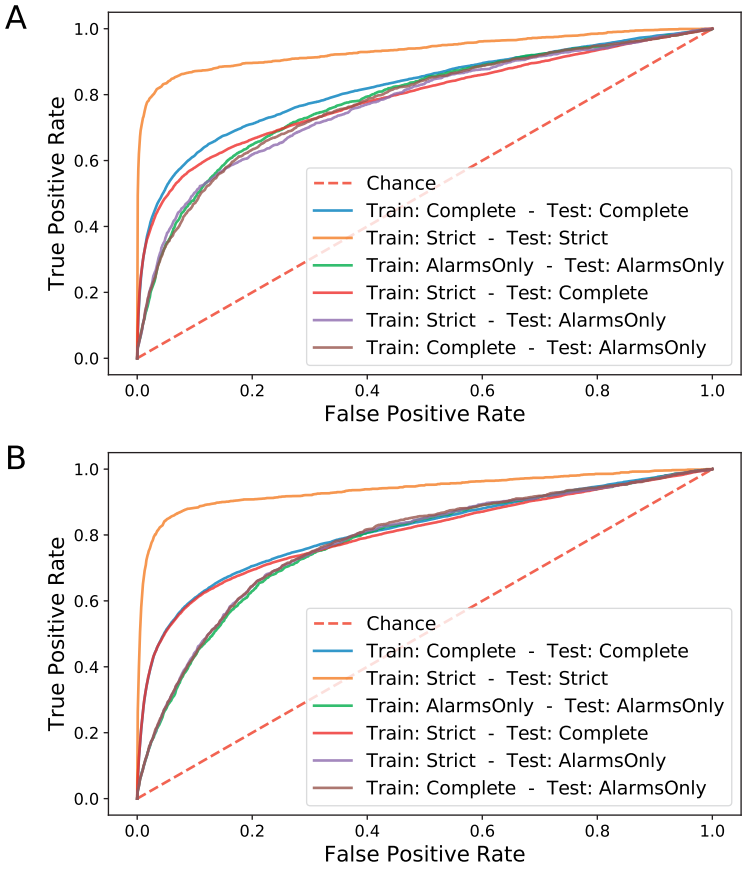


Figure 3.4: Receiver Operating Characteristic curves (ROC) for the detection of caregiver response for all experiments computed by means of (A) logistic regressor and (B) nearest mean classifier.

Results from this table show that the highest AUROC value was found by considering the strict labelling for both the training and testing dataset and the use of the nearest mean (experiment 2). This result allowed to associate a high true positive rate (0.90), main priority in a NICU due to complications associated to missed responses to critical conditions, with a very low false positive rate (0.20), indicating a very low incidence of unnecessary requested responses, a result significantly better compared to the high number of false or irrelevant alarms often reported in studies related to NICU environments [9], [10]. This result indicated that both the logistic regressor and the nearest mean classifier can learn the difference between a stable and a deteriorating patient.

High AUROC values were also found when the complete or strict labelling were used for the training and complete labelling was used for the testing (experiment 1 and 4, respectively). **Table 3.5** presents the confusion matrices for one of the testing folds selected at random for the nearest mean classifier for these experiments. These were calculated for a threshold in the best operational point equal to 0.65 extracted from the ROC curves. Confusion matrices were further split considering labelling for alarms and responses as defined in **Paragraph 3.5.3**. It is also important to notice that the confusion matrices for experiments 5 (training with strict, testing with AlarmsOnly) and 6 (training with complete, testing with AlarmsOnly) can also be derived from the confusion matrices of experiment 4 and 1, respectively. This is because the number of windows used for testing depends on the AlarmOnly labelling, which is a subset of the windows used for complete labeling, and because these pairs of experiments shared the same labelling scheme for the training.

Experiment 1 showed that we can detect all responded events with a mean AUROC value equal to 0.82 for the logistic regressor, without much variance between the subject folds. With experiment 4 instead, the use of the strict label while training for the same goal caused a slight decrease in the AUROC performance. Moreover, while the precision and specificity increased for the fold presented in **Table 3.5** from experiment 1 (precision = 0.14, specificity = 0.94) to experiment 4 (precision = 0.29, specificity = 0.98), the sensitivity (recall) decreased in experiment 4 (0.34) compared to experiment 1 (0.54).

We speculated that by training with complete labelling the classifiers might also discern other types of interactions and cause therefore a decrease in precision, whereas by training with the strict labelling the focus can be moved on deterioration only. This hypothesis was also supported by the original labels for each experiment, also presented in **Table 3.5**. In the decomposed confusion matrix for experiment 1 it is shown that, out of the 314 false negatives, only 35 were deterioration events (i.e., AR = 11). Most of the remaining 279 false negatives could potentially be linked to routine interactions, not entailing therefore a clinical deterioration (e.g., caregiver handling). During routine interactions a response (i.e., interaction with the patient) is in fact present but the patient vital signs are stable during the interaction, events that might therefore lead the classifier to a wrong decision. The false negatives for responded events (i.e., AR = 01) further increased with experiment 4. In addition to this, while the true positives for deterioration events (i.e., AR = 11) also reduced from experiment 1 to 4 by about 25% (153 vs. 114), this reduction was around 50% for responded events (218 vs. 118). Unfortunately, there were no records in the dataset that explicitly showed the type of interaction (e.g., deterioration, caregiver handling, parents) so we can not determine a fully correct evaluation per each window. Yet, this knowledge could help explaining

several classification errors and gave us relevant insides for future extensions of this work.

Table 3.5: Confusion matrices and their respective true classes for a random fold and experiment 1 (training with complete, testing with complete) (A) and experiment 4 (training with strict, testing with complete) using the nearest mean classifier and a threshold equal to 0.65 (best operational point in ROC).

A	Estimated labels		
True labels	Not resp.	Resp	Total
Not resp.	33218	2297	35515
Resp	314	371	685

True labels	Estimated labels		Total
AR = 00	31530	1401	32931
AR = 10	1688	896	2584
AR = 01	279	218	497
AR = 11	35	153	188

B	Estimated labels		
True labels	Not Resp.	Resp	Total
Not resp.	34944	571	35515
Resp	453	232	685

True labels	Estimated labels		Total
AR = 00	32673	258	32931
AR = 10	2271	313	2584
AR = 01	379	118	497
AR = 11	74	114	188

No particular changes when found by comparing the ROC and AUROC values for experiment 3, 5 and 6, as shown in **Table 3.4** and **Figure 3.4**. Furthermore, as previously mentioned, the AlarmOnly labelling (used during testing) used a subset of windows compared to the complete labelling. Metrics extracted for one fold from experiment 1 corresponded exactly to those obtained for experiment 6 and those extracted from experiment 4 corresponded to those obtained for experiment 5, respectively, and can therefore be extracted from **Table 3.5**. By considering only windows with alarms and no response (i.e., AR = 10) and deterioration events (i.e., AR = 11) in the confusion

matrices, the precision and specificity increased for the fold presented in **Table 3.5** from experiment 1 (precision = 0.15, specificity = 0.65) to experiment 4 (precision = 0.27, specificity = 0.88) while the sensitivity decreased in experiment 4 (0.61) compared to experiment 1 (0.81). This implied that by training with the strict labelling (experiments 4 and 5) it is possible to obtain an increase in precision without drastically compromising the true positives. This unfortunately came with a reduction in the sensitivity that could also be explained by the classifier having a stricter concept for a deteriorating event.

3.6.2. Time analysis of probabilities

In addition to the global performance, we present the time analysis of the test data for 24 hours of one patient selected at random using windows of 1 minute. **Figure 3.5** presents the posterior probability of caregiver response when training with the complete labelling (top) and with the strict labelling (middle). True windows labelling based on the presence or absence of alarms (A) and responses (R) as presented in **Paragraph 3.5.3** is also shown (bottom).

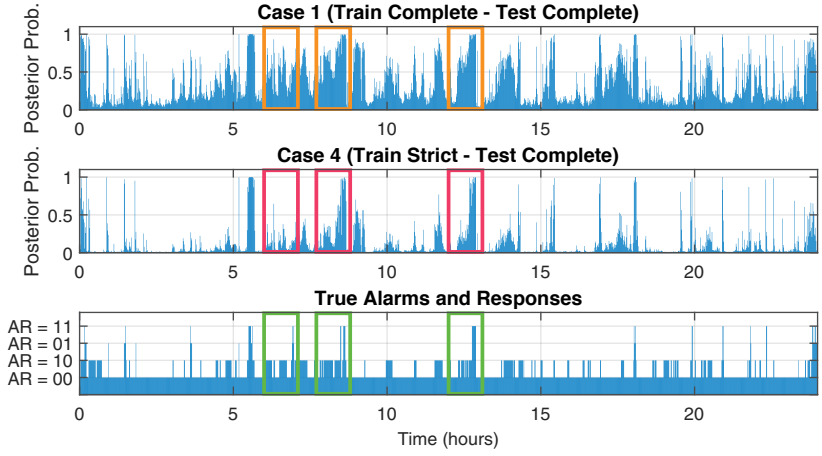


Figure 3.5: Time analysis of probabilities of the logistic regressor for one day of samples of a random test subject. (Top) Experiment 1: training with complete, testing with complete. (Middle) Experiment 4: training with strict, testing with complete. (Bottom) True windows labelling based on the presence of absence of alarms (A) and responses (R). Colored boxes are also used to indicate probabilities around hours in which the presence of alarms with responses (AR = 11) and no responses (AR = 10) were present and around which experiment 1 and 4 performed quite differently.

When training with the complete labelling (i.e., experiment 1), several high probabilities were found around times in which alarms were triggered by the monitor but there was no response yet (AR = 10), as it is for instance highlighted with colored boxes for

hour 6, 8 and 13. In contrast, when training with the strict label, which only considers true deterioration events, the classifier was capable of reducing the probability for windows with only alarms while maintaining high probabilities for sections around true deteriorations (AR = 11).

This analysis helped showing that in a timely manner the classifier trained with complete labelling was able to point to deterioration events in a general way but had problems choosing the exact windows that overlapped with the ground truth. This mismatch could also explain the low precision values derived from **Table 3.4**, as these low performances were mostly due to mismatches in windows around the deterioration events. On the other hand, as can be seen in **Figure 3.5**, the classifier for experiment 4 (i.e., training with strict labelling) triggered a high probability around almost all deterioration events present in this day.

We hypothesize that different window sizes or adapting the evaluation to consider this time dependency could potentially show fairer global results for precision. But we leave this for future efforts as it lies outside of the scope of this chapter.

3.7. Conclusions and future works

In this chapter we introduced our method to detect moments that need to be responded to by caregivers in a NICU, using multiple modalities extracted from signals in bedside patient monitors and caregiver response to alarms. This response comprises an inherently learned knowledge by the medical team of interpreting the nonverbal cues displayed by the patients, which combined with the information from monitoring signals better informs the medical team of the true state of the patients and their potential deteriorations.

We obtained a mean AUROC value equal to 0.82 for all responded events when using only information regarding caregiver responses (i.e., experiment 1), a relevant result considering the high percentages of false alarms currently sounding in NICUs and causing alarm fatigue in caregivers.

We used different labeling schemes for caregiver response, motivated in the type of interaction or response present in our data, both from a medical and a social point of view (e.g., kangaroo care). We showed that while training only considering responses to detect responded events (i.e., experiment 1) returned a better general performance, whereas training only with stable windows and windows with deterioration to detect

responded events (i.e., experiment 4) helped reducing the number of alarms triggered by the patient monitor with no need of response.

We indicated that a better evaluation of types of interaction between caregiver and patients (e.g., further analyzing nonverbal cues displayed by the patients and their correlations with their health condition) can further help reducing our number of false positives. Additionally, incorporating time dependency in our algorithms could potentially improve the detection of responded events.

This study provides a first insight into the possibility of defining new meaningful alarms based on previously collected caregiver responses, as a prior for a more comprehensive understanding of the state of the patients that includes interpreting their nonverbal cues. Together with these results, future studies may shed a light on the possibility of revisiting alarms used in the NICU, by replacing threshold-based alarms with alarms defined based on previous caregiver responses, and possibly have an impact in reducing alarm fatigue.

References

[1] S. Siebig, S. Kuhls, M. Imhoff, U. Gather, J. Schölmerich, and C. E. Wrede, "Intensive care unit alarms-How many do we need?," *Crit. Care Med.*, vol. 38, no. 2, pp. 451–456, 2010.

[2] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomed Instrum Technol.*, pp. 46 (4), 268–277, 2012.

[3] S. Sendelbach and M. Funk, "Alarm fatigue: A patient safety concern," *AACN Adv. Crit. Care*, vol. 24, no. 4, pp. 378–386, 2013.

[4] B. M. Funk, J. T. Clark, T. J. Bauld, J. C. Ott, and P. Coss, "Attitudes and Practices Related to Clinical Alarms," *Am. J. Crit. Care*, vol. 23, no. 3, pp. e9–e18, 2014.

[5] C. Van Pul *et al.*, "Alarm management in a single-patient room intensive care units," *Recent Adv. Ambient Assist. Living – Bridg. Assist. Technol. e-Health Pers. Heal. Care*, pp. 119–132, 2015.

[6] C. Van Pul, H. Van De Mortel, J. Bogaart, T. Mohns, and P. Andriessen, "Safe patient monitoring is challenging but still feasible in a neonatal intensive care unit with single family rooms," *Acta Paediatr. Int. J. Paediatr.*, vol. 104, no. 6, pp. e247–e254, 2015.

[7] W. Chen, S. B. Oetomo, and L. Feijs, "Neonatal monitoring: Current practice and future trends.," in *Handbook of Research on Developments in E-Health and Telemedicine*, 2010, pp. 939–961.

[8] F. Dursun Ergezen and E. Kol, "Nurses' responses to monitor alarms in an intensive care unit: An observational study," *Intensive Crit. Care Nurs.*, vol. 59, p. 102845, 2020.

[9] R. Joshi, H. Van De Mortel, L. Feijs, P. Andriessen, and C. Van Pul, "The heuristics of nurse responsiveness to critical patient monitor and ventilator alarms in a private room neonatal intensive care unit," *PLoS One*, vol. 12, no. 10, pp. 1–14, 2017.

[10] B. J. Drew *et al.*, "Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients," *PLoS One*, vol. 9, no. 10, 2014.

[11] D. Ostojic *et al.*, "Reducing false alarm rates in neonatal intensive care: A new machine learning approach," *Adv. Exp. Med. Biol.*, vol. 1232, no. January, pp. 285–290, 2020.

[12] J. P. Keller, "Clinical alarm hazards: A top ten health technology safety concern," *J. Electrocardiol.*, vol. 45, no. 6, pp. 588–591, 2012.

[13] H. Ruppel, M. Funk, R. Whittemore, S. F. Wung, C. P. Bonafide, and H. Powell Kennedy, "Critical care nurses' clinical reasoning about physiologic monitor alarm customisation: An interpretive descriptive study," *J. Clin. Nurs.*, vol. 28, no. 15–16, pp. 3033–3041, 2019.

[14] C. McClure, S. Y. Jang, and K. Fairchild, "Alarms, oxygen saturations, and SpO2 averaging time in the NICU," *J. Neonatal. Perinatal. Med.*, vol. 9, no. 4, pp. 357–362, 2016.

[15] B. J. Stenson *et al.*, "Oxygen saturation and outcomes in preterm infants," *N. Engl. J. Med.*, vol. 368, no. 22, pp. 2094–2104, 2013.

[16] J. Welch, "An evidence-based approach to reduce nuisance alarms and alarm fatigue," *Biomed. Instrum. Technol.*, vol. 45, no. SPRING, pp. 46–52, 2011.

[17] G. Varisco *et al.*, "Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU," *Acta Paediatr. Int. J. Paediatr.*, vol. 110, no. 4, pp. 1141–1150, 2021.

[18] L. Chen *et al.*, "Using supervised machine learning to classify real alerts and artifact in online multisignal vital sign monitoring data," *Crit. Care Med.*, vol. 44, no. 7, pp. e456–e463, 2016.

[19] L. M. Eerikainen, J. Vanschoren, M. J. Rooijakkers, R. Vullings, and R. M. Aarts, "Reduction of false arrhythmia alarms using signal selection and machine learning," *Physiol. Meas.*, vol. 37, no. 8, pp. 1204–1216, 2016.

[20] A. H. Gee, R. Barbieri, D. Paydarfar, and P. Indic, "Predicting Bradycardia in Preterm Infants Using Point Process Analysis of Heart Rate," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2300–2308, 2017.

[21] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: A review," *Med. Biol. Eng. Comput.*, vol. 44, no. 12, pp. 1031–1051, 2006.

[22] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, "Heart rate variability Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, no. 17, pp. 354–381, 1996.

[23] J. R. Williamson, D. W. Bliss, D. W. Browne, P. Indic, E. Bloch-Salisbury, and D. Paydarfar, "Using physiological signals to predict apnea in preterm infants," *Conf. Rec. - Asilomar Conf. Signals, Syst. Comput.*, pp. 1098–1102, 2011.

[24] C. J. Alonzo, V. P. Nagraj, J. V Zschaebitz, D. E. Lake, J. Randall Moorman, and M. C. Spaeder, "Heart rate ranges in premature neonates using high resolution physiologic data," *J Perinatol.*, vol. 38, no. 9, pp. 1242–1245, 2018.

[25] V. Golder, M. Hepponstall, S. R. Yiallourou, A. Odoi, and R. S. C. Horne, "Autonomic cardiovascular control in hypotensive critically ill preterm infants is impaired during the first days of life," *Early Hum. Dev.*, vol. 89, no. 6, pp. 419–423, 2013.

[26] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Min.*, vol. 10, no. 1, pp. 1–17, 2017.

[27] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1260–1284, 2019.

CHAPTER 4

Music therapy improves autonomic regulation in premature infants

Reproduced after: Effect of Music Therapy Interventions on Heart Rate Variability in Premature Infants, 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 678-681.

Abstract

Premature infants are at risk of developing serious complications after birth. Communicative interventions performed in neonatal intensive care units (NICUs), such as music therapy interventions, can reduce the stress experienced by these infants and promote the development of their autonomic nervous system. In this study we investigated the effects of music therapy interventions, consisting of singing, humming, talking or rhythmic reading, on premature infants by investigating the effects on their heart rate variability (HRV).

A total of 27 communicative interventions from 18 patients were included in this study. The NN-intervals were extracted from the ECG and the mean ± SEM values were investigated for 9 different features (HR, SDNN, RMSSD, pNN50, pDec, sDDec, LF Power, HF Power and LF/HF Ratio). Median feature values for the pre- and communicative intervention were compared using the Wilcoxon signed-rank test.

An increase in values for the SDNN, RMSSD and pNN50 was found in the 20 min preceding the communicative interventions, when caregiving activities were performed, and was followed by an immediate decrease at the start of the communicative interventions. Features' variability during the communicative interventions appeared to be smaller than in the pre-communicative interventions, indicating improved autonomic regulation. Significant differences ($P < 0.05$) were found for the LF Power and HF Power, with lower values found during the communicative interventions, results that, in accordance with previous studies, are associated with stress reduction (i.e., reduced sympathetic dominance). Communicative interventions can therefore be useful for stabilizing the autonomic nervous system in premature infants.

4.1. Introduction

Prematurity increases the risk of developmental problems and long-term comorbidity. Most infants born before 30 weeks of gestational age are admitted to specialized neonatal intensive care units (NICU), where care is aimed at stabilizing physiological parameters such as heart rate (HR), oxygen uptake and sleep to promote growth. However, a NICU is considered an unnatural environment for the premature infant, compared to the safe environment inside the mother's womb. This results in stress due to the separation from the parent, but also because of the bright lights, noise, several interventions and other unnatural stimuli [1]–[3].

Parent-infant interactive interventions can reduce the stress of the infants. Feldman et al. indicated that skin-to-skin contact between the premature infant and the parents aids in the development of the autonomic nervous system [4]. The vagal tone develops more rapidly for premature infants between 32 and 37 weeks gestational age if there is skin-to-skin contact [5]. Additionally, music therapy has also shown beneficial effects on the autonomic nervous system, leading to increased brain connectivity [6]. It has also been reported that the maternal voice improves the physiological regulation of the premature infant, with an increase in the oxygen saturation of the infant, a decrease in HR and a decrease in number of cardiorespiratory events [7]–[10].

Additional information about regulatory changes in premature infants can be provided by the heart rate variability (HRV). HRV is a measure of the variation in time between consecutive heartbeats. Acceleration and deceleration of the HR are regulated by the two complementary subsystems of the autonomic nervous system, the parasympathetic and sympathetic nervous system. As such, changes in these systems will impact the HRV through changes in the autonomic regulation [11]. HRV is one of the most direct and easily accessible features that give information about the autonomic nervous system maturation of the infant [12]. Frequency analysis of HRV, extracted from photoplethysmography signals, showed effects in the low and high frequency power ranges, interpreted as enhanced vagal activity in these infants during hospitalization [9]. In our study, we explored the effect of communicative interventions on premature infants in HRV-features calculated from ECG-signals. The communicative interventions consisted of singing, humming, talking or rhythmic reading and were guided by a music therapist.

4.2. Methods

4.2.1. Participants and communicative interventions

Communicative interventions were offered to all parents of clinically stable premature infants with a gestational age > 30 weeks in the neonatology units of Máxima Medical Center (both NICU and the neonatal medium care unit). The communicative intervention consisted of approximately 20 to 30 min in which the music therapist stimulated the parents to make contact with their infant by talking, singing, humming or rhythmic reading. These actions were accompanied by guitar play by the music therapist. When the music therapist observed changes in the reaction or the behavior of the infant, the parents were further guided in their interaction to promote rhythmic breathing of the infant and parent-infant attachment.

To prevent the interruption of the regular clinical practice, all the caregiving activities were performed before and after the communicative intervention. During the communicative intervention, the infant was placed lying in an incubator, in an open bed or in a prone position on the parent's chest. It was considered important for the premature infant to be awake or in active sleep during the communicative intervention. If the infant was crying or in a deep sleep condition, the communicative intervention was postponed. The timeline of the communicative intervention is displayed in **Figure 4.1**.

This study, being an observational study in line with the clinical practice, received a waiver from the medical ethical committee of the Máxima Medical Center, in accordance with the Dutch law on medical research with humans (WMO). Written informed consent was obtained from the parents to allow for the data collection.

One or both parents were always present in the room during the communicative intervention. A total of 26 families with 28 premature infants participated in this study. However, since for 10 patients the ECG data was not collected for a full period of 1 h in pre-communicative intervention, during the communicative intervention or for a full period of 1 h in post-communicative intervention, these patients were excluded from this analysis.

4.2.2. Data analysis

The 3-lead ECG was acquired from the patient monitoring system (IntelliVue MX 800, Philips, Hamburg, Germany) with a sampling frequency of 250 Hz. The monitored signals were saved automatically in a private data warehouse (PIIC iX, Data Warehouse Connect, Philips Medical Systems, Andover, MA). The ECG signal was analyzed 1 h before, during and 1 h after the communicative intervention. As the duration of the communicative

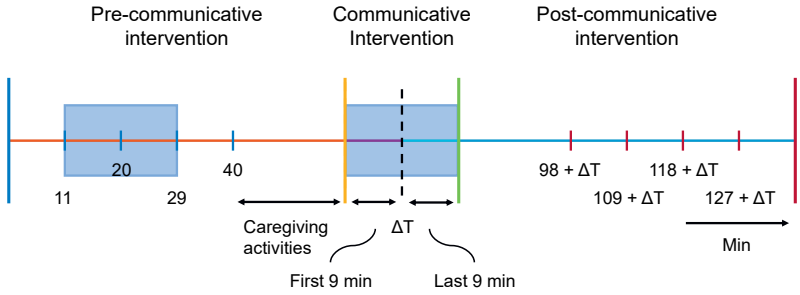


Figure 4.1: Timeline of the communicative intervention. As the duration of the communicative intervention was variable, the first 9 and the last 9 min were used for analysis. ΔT represents the time between the first and last 9 min of the communicative intervention. The blue blocks represent the time frames used for the statistical analysis.

intervention was variable, the first 9 min and the last 9 min of the communicative intervention were selected to be included in an 18-minute analysis.

During the pre-communicative intervention, the premature infant had caregiving activities such as feeding and diaper change, mainly in the last 20 min of this period (indicated as 'patient handling period' in **Figure 4.1**).

The ECG recordings were analyzed using Matlab version R2020a (The MathWorks, Natick, USA). A peak detection algorithm was used to identify the R-peaks in the ECG recordings. Consequently, the time distance between successive normal heart-beats R-peaks (e.g., not ectopic) was computed to determine the NN-intervals.

Features were then extracted from the NN-intervals considering 1-minute-long time windows, unless ECG data was found to be missing for more than 10 s within these time windows. A total of 9 features in the time-domain and frequency-domain were included in this study. The first group included: 1) the HR; 2) the SDNN, standard deviation of the NN-intervals [13]–[15]; 3) the RMSSD, root mean square of successive differences between normal R-peaks [13]–[15]; 4) the pNN50, percentage of absolute differences in successive NN-intervals exceeding 50 ms [11], [14], [16]; 5) the pDec, percentage of NN-intervals that arise from decelerations and exceeds the mean NN-interval of the previous 5 min [11], [17] and 6) the sDDec, standard deviation of the HR decelerations [11], [17]. In order to compute features in the frequency-domain, the NN-intervals were resampled using a cubic spline interpolation and a sampling frequency of 250 Hz, similar to the original sampling frequency of the ECG. The Fast Fourier Transform (FFT) was then used to extract 7) the LF Power, measure of the signal energy within the low frequency range (0.02-0.2 Hz), 8) the HF Power, measure of the signal energy within the high frequency range (0.2-1.5 Hz) and 9) the LF/HF Ratio, ratio between the LF Power and HF Power,

defined following the indication of previous studies [9], [18]. For all the HRV features the mean ± SEM values were calculated for each minute. For the comparison between infants, baseline removal was performed by subtracting the mean value of the last 20 min of pre-communicative intervention from the mean value of each HRV feature.

For the statistical analysis, median values of the HRV features for the communicative intervention were compared to a stable time frame in the pre-communicative intervention, as indicated with two blue boxes in **Figure 4.1**, by using the Wilcoxon signed-rank test.

4.3. Results

A total of 74 communicative interventions were given across 26 families (28 infants). The minimum number of communicative interventions was one, the maximum ten. After checking the available ECG data and therapy duration, 27 communicative interventions were included from 11 infants who had one communicative intervention, 5 infants with two and 2 infants with three.

The mean ± SEM values for the HR, SDNN, pNN50 and LF Power are shown as a function of time in **Figure 4.2**, where the dotted lines separate the pre-communicative intervention, the first and last 9 min of communicative intervention and the post-communicative intervention. Note that only 4 out of 9 features were included since other features showed similar trends.

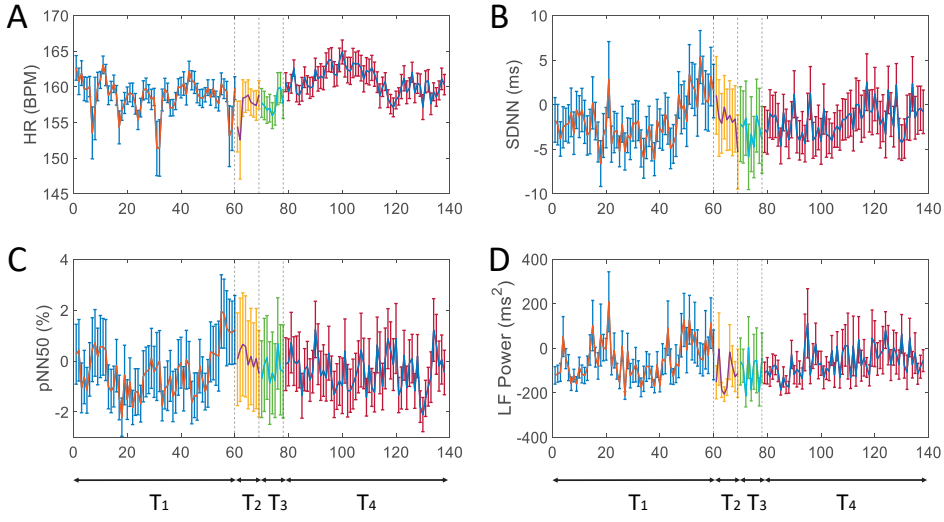


Figure 4.2: Time-series for the (A) HR, (B) SDNN, (C) pNN50 and (D) LF Power. Mean ± SEM values for all patients after baseline removal are shown for the pre-communicative intervention (T1), the first 9 min of communicative intervention (T2), the last 9 min of communicative intervention (T3) and the post-communicative intervention (T4).

It was observed that in the 20 min preceding the communicative intervention, values for different features increased during patient handling but started to decrease immediately at the start of the communicative intervention. This behavior was found for instance for the RMSSD, LF Power and LF/HF Ratio, and was even more pronounced for the SDNN and pNN50. Variability of features, as shown by the SEM, was found to be generally large. However, smaller SEM were found during the communicative intervention than in the pre- or post-communicative intervention, as it was noticed for the SDNN and LF/HF Ratio.

For all features, the median (IQR) values are shown in **Table 4.1**. Significant differences ($P < 0.05$) were found for the LF and HF Power, with lower values found during the communicative intervention. For all other features, although different values were observed between pre- and communicative intervention, the IQR was found to be large.

4.4. Discussion

Infant-directed communication (like speech or singing) is an auditory form of co-regulation and could have a beneficial effect on the autonomic regulation [19]. In our study, we observed a decreasing trend in different time-domain HRV features (e.g., SDNN, pNN50 and LF/HF Ratio) during the communicative intervention compared to the pre-communicative intervention. Contrastingly, in adults, a substantial decrease in these HRV values in rest is associated with a higher risk of cardiovascular mortality and morbidity, whereas a large SDNN in rest reflects a healthy, rapid, proactive autonomic nervous system regulation with overall parasympathetic dominance [20].

Table 4.1: Median values (with IQR) for HR and HRV features for the pre-communicative intervention (from the 11th and up to the 29th min) and communicative intervention (first and last 9 min) with indication for increase or decrease (+/-) compared to the pre-communicative intervention, without baseline removal. Significance (i.e., p-value < 0.05) is reported with an * whenever it is found.

HR and HRV features	Pre-communicative intervention	Communicative intervention	Significance
HR (bpm)	160 (152-163)	157 (142-164) (-)	
SDNN (ms)	17.5 (12.6-30.5)	15.6 (11.4-23.7) (-)	
RMSSD (ms)	8.6 (5.7-17.1)	9.3 (5.4-16.2) (+)	
pNN50 (%)	0.6 (0.0-5.1)	0.3 (0.0-1.6) (-)	
pDec (%)	49.3 (43.0-57.8)	47.4 (42.8-51.4) (-)	
sDDec (ms)	11.6 (7.1-21.1)	13.4 (7.5-17.5) (+)	
LF Power (ms^2)	145.8 (51.8-396.3)	86.5 (29.4-244.1) (-)	*
HF Power (ms^2)	23.0 (5.2-186.7)	8.2 (3.6-47.9) (-)	*
LF/HF Ratio (%)	5.4 (1.1-14.2)	7.2 (2.2-14.6) (+)	

However, there is a difference in the cardio-respiratory physiology between adults and premature infants. In premature infants too rapid, undesirable decreases in HR occur regularly and an increase in HRV is often associated with instability in these patients, such as during the transfers from and to the incubator [11]. This likely happens in premature infants because the immature autonomic nervous system overshoots to regulate the HR. In fact, in premature infants stress often leads to (severe) bradycardia illustrating the shortcoming of the immature autonomic nervous system to generate optimal responses to cope with environmental challenges [11]. This was reflected in this study by an increase in values for all conventional time-domain HRV features in the last 20 min of pre-communicative intervention where caregiving activities took place. Therefore, the assumption that higher values of HRV are associated with instability of the premature infants is also confirmed in this study. Values for these same features decreased immediately at the start of the communicative intervention, indicating improved autonomic regulation. In addition to this, a small decrease in HR was visible during the communicative intervention compared to the pre-communicative intervention. Clinically a 3 bpm drop in HR may not be relevant, but it does indicate a change in autonomic regulation.

Frequency-domain HRV features showed significant differences when the pre-communicative and communicative intervention were compared, with lower values found for the LF and HF spectral components during the communicative intervention. The interpretation of the LF and HF Power for premature infants is currently still debated and not directly related with adult interpretation for HRV frequency domain features. In previous studies, for instance, an increase in LF Power was found during active waking states in premature infants. It was associated to an increase in activity of the autonomic nervous system altogether and was consequently used as an index of physiological arousal, not stress per se [9], [21]. At the same time an increase in HF Power in premature infants was related to an increase in the parasympathetic activity in some studies [9], [21], suggesting that a decrease in HF power would reflect stress. However, a decrease in HF power was seen in infants during kangaroo care [11]. Nevertheless, a decrease in LF Power greater than a decrease in HF Power has mostly been related to stress reduction [9], [21]. This is in line with the findings in the current study. Here results showed a minimum LF/HF Ratio in the first minutes of the pre-communicative intervention, increasing gradually when caregiving activities were performed and immediately dropping at the start of the communicative intervention.

This study has some important limitations. One limitation is that the therapeutic approach of the communicative intervention was not standardized but instead tailored to the individual development phase of the infant and to the wishes of the parents. Different types of therapy were in fact applied (singing, humming, talking or

rhythmic reading). Observations during the communicative intervention in our study showed that infant responded to the mother's voice (without any musical guidance), for instance by laughing or turning the head towards the mother. However, in our study, due to our choice to provide tailored therapeutic approaches for the different communicative interventions it was not possible to determine whether the different methods could have had a different effect on the HRV, a choice which constituted a first limitation of this study. Future studies should consider the inclusion of a higher number of communicative interventions to perform an individual analysis on each possible type of therapy.

Another limitation of this study was the inhomogeneity of the group, which possibly led to the large variability found for the different HRV features. The infants participating in the communicative interventions had different gestational and postmenstrual ages. It is known that the HRV values can vary between different gestational ages [22]. Another factor that complicates the interpretation of the data was that all infants were allowed to participate in this study, under the condition that the infants were stable at the time of the communicative intervention. Therefore, also infants who suffered from underlying diseases could have participated in the communicative intervention, possibly leading to more variability to the data and causing some effects to go unnoticed.

The last factor that influenced the HRV values was the number of communicative interventions that the infants participated in. In the therapist's notes it was stated that some infants had to get used to the communicative intervention and the more communicative interventions the infant participated in, the better the response was to the communicative intervention. A clear decrease in HR was seen for these infants on the patient monitor along the communicative interventions.

4.5. Conclusion

Changes in HRV-based features were observed during communicative intervention, indicating improved autonomic regulation for the premature infant during the communicative interventions. Frequency-domain features showed a significant decrease in the activity of the sympathetic nervous system during the communicative interventions compared to the instances in which caregiving activities were performed. Communicative interventions can therefore be useful for stabilizing the autonomic nervous system. However, different types of activities were applied in our study (e.g., singing, humming, talking or rhythmic reading) which could potentially have had different influences on the HRV and possibly concealed some effects of the communicative interventions.

References

[1] Z. Marcos, "Arriving too early," *Lancet Neurol.*, vol. 12, no. 4, pp. 332–333, 2013.

[2] K. C. Sekar, "Iatrogenic complications in the neonatal intensive care unit," *J. Perinatol.*, vol. 30, no. June 2009, pp. S51–S56, 2010.

[3] D. I. Maroney, "Recognizing the potential effect of stress and trauma on premature infants in the NICU: How are outcomes affected?," *J. Perinatol.*, vol. 23, no. 8, pp. 679–683, 2003.

[4] R. Feldman and A. I. Eidelman, "Skin-to-skin contact (Kangaroo Care) accelerates autonomic and neurobehavioural maturation in preterm infants," *Dev. Med. Child Neurol.*, vol. 45, no. 4, pp. 274–281, 2003.

[5] M. I. Porges, Stephen W Davila *et al.*, "Autonomic regulation of preterm infants is enhanced by Family Nurture Intervention," *Dev Psychobiol.*, vol. 61, no. 6, pp. 942–952, 2019.

[6] F. B. Haslbeck, A. Jakab, U. Held, D. Bassler, H. U. Bucher, and C. Hagmann, "Creative music therapy to promote brain function and brain structure in preterm infants: A randomized controlled pilot study," *NeuroImage Clin.*, vol. 25, no. October 2019, p. 102171, 2020.

[7] L. Provenzi, S. Broso, and R. Montirosso, "Do mothers sound good? A systematic review of the effects of maternal voice exposure on preterm infants' development," *Neurosci. Biobehav. Rev.*, vol. 88, no. March 2017, pp. 42–50, 2018.

[8] M. Filippa *et al.*, "Systematic review of maternal voice interventions demonstrates increased stability in preterm infants," *Acta Paediatr.*, vol. 106, no. 8, pp. 1220–1229, 2017.

[9] M. Filippa *et al.*, "Maternal Singing but Not Speech Enhances Vagal Activity in Preterm Infants during Hospitalization: Preliminary Results," *Children*, vol. 9, no. 2, pp. 1–10, 2022.

[10] S. M. Ludington-Hoe, K. Morgan, and A. Abouelfettoh, "A Clinical Guideline for Implementation of Kangaroo Care With Premature Infants of 30 or More Weeks' Postmenstrual Age," *Adv. Neonatal Care*, vol. 8, no. 3, pp. S3–S23, 2008.

[11] D. R. Kommers *et al.*, "Features of Heart Rate Variability Capture Regulatory Changes During Kangaroo Care in Preterm Infants," *J. Pediatr.*, vol. 182, pp. 92-98.e1, 2017.

[12] R. Sahni, K. F. Schulze, S. Kashyap, K. Ohira-Kist, F. W. P., and M. M. Myers, "Maturational changes in heart rate and heart rate variability in low birth weight infants," *Dev Psychobiol.*, vol. 37, no. 2, pp. 73–81, 2000.

[13] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, "Heart rate variability Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, no. 17, pp. 354–381, 1996.

[14] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Front. Public Heal.*, vol. 5, no. September, pp. 1–17, 2017.

[15] H. Nagendra, V. Kumar, and S. Mukherjee, "Cognitive behavior evaluation based on physiological parameters among young healthy subjects with yoga as intervention," *Comput. Math. Methods Med.*, vol. 2015, 2015.

[16] J. E. Mietus, C. K. Peng, I. Henry, R. L. Goldsmith, and A. L. Goldberger, "The pNNx files: Re-examining a widely used heart rate variability measure," *Heart*, vol. 88, no. 4, pp. 378–380, 2002.

[17] L. Cabrera-Quiros *et al.*, "Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning," *Crit. Care Explor.*, vol. 3, no. 1, p. e0302, 2021.

[18] E. G. Rosenstock, Y. Cassuto, and E. Zmora, "Heart Rate Variability in the Neonate and Infant: Analytical Methods, Physiological and Clinical Observations," *Acta Paediatr*, vol. 88, no. 5, pp. 477–82, 1999.

[19] M. Van Puyvelde and F. Franco, "The interaction of music and language in the ontogenesis of human communication: A multimodal parent-infant co-regulation system," *Middlesex Univ. Res. Repos.*, no. April 2016, pp. 1–8, 2015.

[20] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: A review," *Med. Biol. Eng. Comput.*, vol. 44, no. 12, pp. 1031–1051, 2006.

[21] K. Javorka *et al.*, "Heart rate variability in newborns," *Physiol. Res.*, vol. 66, no. 2, pp. S203–S214, 2017.

[22] E. Longin, T. Gerstner, T. Schaible, T. Lenz, and S. König, "Maturation of the autonomic nervous system: Differences in heart rate variability in premature vs. term infants," *J. Perinat. Med.*, vol. 34, no. 4, pp. 303–308, 2006.

PART TWO

New insights into central apnea and its detection

CHAPTER 5

Facilitating annotations for central apnea in premature infants

Reproduced after: Optimized Detection of Central Apneas Preceding Late-Onset Sepsis in Premature Infants, 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 5463-5468.

Abstract

In neonatal intensive care units, respiratory traces of premature infants developing late onset sepsis (LOS) may also show episodes of apneas. However, since clinical patient monitors often underdetect apneas, clinical experts are required to investigate patients' traces looking for these events. In this work we present a method to optimize an existing algorithm for central apnea (CA) detection and how we used it together with human annotations to investigate the occurrence of CAs preceding LOS.

The algorithm was optimized by using a previously annotated dataset consisting of 90 hours, extracted from 10 premature infants. This allowed to double precision (19.7% vs 9.3%, median values per patient) without affecting recall (90.5% vs 94.5%) compared to the original algorithm. This choice caused the missed identification of just 1 additional CA (4 vs 3) in the whole dataset. The optimized algorithm was then used to annotate a second dataset consisting of 480 hours, extracted from 10 premature infants diagnosed with LOS. Annotations were corrected by two clinical experts.

A significantly higher number of CA annotations was found in the 6 hours prior to sepsis onset (p-value < 0.05). The use of the optimized algorithm followed by human annotations proved to be a suitable, time-efficient method to annotate CAs before sepsis in premature infants, enabling future use in large datasets.

5.1. Introduction

Infants born very prematurely are characterized by an immature respiratory system upon delivery. As a consequence, these infants require hospitalization in neonatal intensive care units (NICUs) [1], [2]. In particular, an immature respiratory system may result in apneic events, or so-called apnea of prematurity [3]–[6]. Previous studies indicated an increase in the number of apneas preceding sepsis, an infection which is among the leading cause of premature death in NICUs [7]–[10]. Therefore, apnea detection is important to adjust the therapeutic strategy to prevent the occurrence of sepsis.

The definition of apnea of prematurity has often been debated in the past due to a lack of consensus among clinical experts [2], [3], [11]. Apneas are currently most often defined as cessations of breathing longer than 20 seconds, or longer than 10 seconds accompanied by a bradycardia (heart rate ≤ 100 beats/minute) and/or desaturation (SpO_2 ≤ 80%) [2], [4]. Apneas can be further classified into different types. These include central apneas (CAs, absence of respiratory effort due to a cessation of output from the central respiratory centers), obstructive apneas (indicated by an obstruction in the upper airway which causes inadequate respiratory efforts to maintain ventilation) and mixed apneas (which include characteristics of both previous groups) [3], [7], [12], [13].

Different algorithms have been developed for automatic detection of apnea (e.g., in patient monitoring systems) [12], [14]–[17]. However, since patient signals show high interpatient and intra-patient variability, such automated apnea detection often fails, particularly in premature infants [12], [17]. Research to improve algorithms for the detection of apnea of prematurity is thus ongoing [15], [16]. For instance, Lee and colleagues proposed an algorithm focusing primarily on CA detection, with a key feature being the removal of the cardiac signal from the chest impedance (CI) as that may falsely be detected as respiration effort [12], [14], [17].

The aim of our study is to facilitate the time-consuming detection of CAs in data tracings of premature infants by optimizing Lee's detection algorithm [12], [14], [17] and to investigate its ability to detect CAs preceding late-onset sepsis (LOS), defined as sepsis with an onset after the first 72 hours of life [9], [10]. This algorithm was selected for the use in this study due to its simplicity in the implementation and since, compared to other algorithms that make use of several additional vital signs to predict sepsis onset [18], [19], it only uses the information included in CI and RR-intervals (the latter is used to filter out the cardiac artifact from CI) for the specific identification of CAs [12].

5.2. Methods

5.2.1. Study design

The improvement of CA detection was addressed in the following two steps, as presented in **Figure 5.1A**. First, Lee's algorithm [12], [14], [17] was optimized using a first dataset (training dataset), by comparing all central apnea-suspected events (CASEs) returned by the algorithm with apnea annotations by two clinical experts. Second, the optimized algorithm was applied in a second dataset (sepsis dataset), containing data from premature infants that developed LOS.

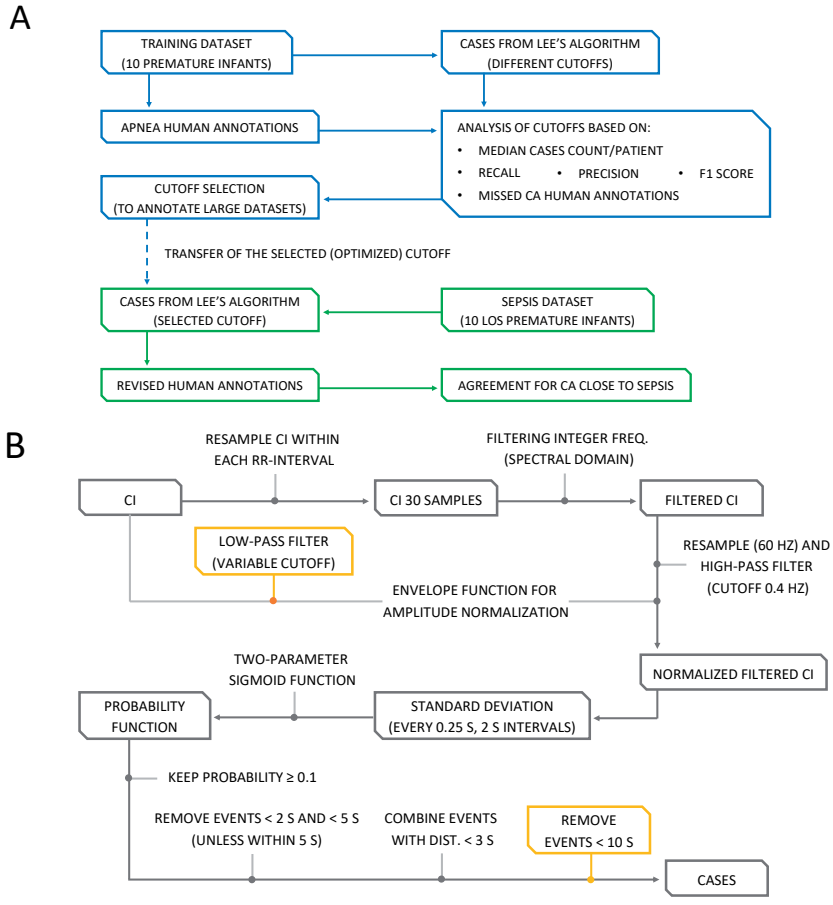


Figure 5.1: (A) Workflow of the steps included in the current study and (B) steps included in Lee's algorithm for central apnea detection. Steps that led to the optimized cutoff for the envelope function of Lee's algorithm are represented in blue in (A), whereas steps to evaluate the agreement for CA before sepsis onset are represented in green. Steps that were optimized or added in this work for Lee's algorithm are highlighted in orange in (B). Abbreviations: CI - chest impedance, CASE - central apnea suspected events, CA – central apnea.

Table 5.1: Characteristics for all patients included in both datasets.

Dataset	Training	Sepsis
Gestational age (weeks)	28.5 ± 1.6	27.9 ± 0.9
Postnatal days	7.0 ± 7.0	10.2 ± 8.1
Birth weight (g)	1041.5 ± 203.4	998.0 ± 215.9
Gender (female)	30%	50%
Hours per patient	9	48
Total hours	90	480

Two clinical experts examined the vital signs during episodes returned as CASEs by the algorithm, and annotated CAs. CASEs as well as matching CAs were evaluated as a function of the distance from sepsis onset.

As this study had a retrospective nature, a waiver was provided by the medical ethical committee in accordance with the Dutch law on medical research with humans (WMO).

5.2.2. Study datasets

Training dataset: This dataset included 10 premature infants (gestational age 28.5 ± 1.6 weeks) undergoing continuous cardiorespiratory monitoring in the NICU of Máxima Medical Center (MMC, Veldhoven, The Netherlands). Patients' characteristics are presented in **Table 5.1**. A total of 90 hours of continuous data was extracted from these patients. ECG, CI and oxygen saturation (SpO_2) were collected from NICU bedside monitors (GE Healthcare, models Solar 8000M and I), with both ECG and CI waveforms tracked using three ECG leads with electrodes on both sides of the heart. ECG was collected at 240 Hz, whereas the CI was collected at 60 Hz. SpO_2 was captured by PPG at 0.5 Hz.

Sepsis dataset: This dataset included 10 premature infants (gestational age 27.9 ± 0.9 weeks) diagnosed with LOS. All patients were part of a sepsis cohort consisting of patients admitted to the NICU of MMC [8]. Patients' characteristics are presented in **Table 5.1**. Their data was collected for all 48 hours prior to the onset of sepsis (when cultures, resuscitation and antibiotics started), which allowed to extract a total of 480 hours of data. ECG, CI and SpO_2 were monitored using the Philips IntelliVue MX800 patients' monitors (Philips Medical Systems, Böblingen, Germany), according to clinical standard. ECG and CI were measured using three ECG leads. ECG was collected at 250 Hz, whereas CI was collected at 62.5 Hz. SpO_2 was captured by PPG at 1 Hz.

5.2.3. Detection of central apnea-suspected events based on Lee's algorithm

Starting point of our work was the central apnea detection algorithm described by Lee [12], [14], [17], that generated a filtered respiration signal without cardiac artifacts. The workflow including all the steps for this algorithm is presented in **Figure 5.1B**. For each RR-interval, the corresponding CI was resampled to provide 30 equidistant samples. The Fourier transform was computed and the frequency band at the integer frequencies was filtered out to remove the cardiac artifact. The signal resulting from inverse transformation was then resampled at 60 Hz and filtered with a high-pass filter with a cutoff frequency of 0.4 Hz, to remove low-frequency artifacts. The amplitude of the resulting signal was finally adjusted using an envelope function. This was computed by applying a low-pass filter with a very low cutoff frequency at the original CI. This was done to normalize the amplitude of the resulting filtered respiration, so that results for different patients could be compared. The value proposed by Lee et al. for this cutoff was 0.0025 Hz. In this study, however, we optimized its value to fit apnea annotations provided by clinical experts from our group for the training dataset.

The standard deviation of the filtered respiration was then computed every quarter of a second for centered 2 second intervals. Apnea probability was defined by applying a two-parameter sigmoid function, resulting in a continuous probability function. Events starting when apnea probability exceeded 0.1 and ending when it decreased below this value were detected. After removing events shorter than 2 seconds, events shorter than 5 seconds were discarded unless they were within 5 seconds of another event. Finally, all events that were separated by less than 3 seconds were combined into one single event, referred as CASEs in this work. Based on apnea definition from [2], [4], only CASEs longer than 10 seconds were considered for further human annotations for CAs.

5.2.4. Central apnea annotations by clinical experts

Training dataset: This dataset had previously been annotated by two clinical experts from MMC. Annotations were performed according to the definition of apnea of prematurity, irrespectively of its type, provided in [2], [4]. The presence of apnea was evaluated using 3 minutes windows from this dataset, visualized by means of an in-house annotation software developed using Matlab (The MathWorks, Natick, USA). The software interface consisted of five physiological signals: ECG, SpO_2, RR-intervals (derived from ECG), CI and respiration rate (derived from CI). All included physiological signals were presented in three minutes segments. An exception was the ECG signal, for which just the last ten seconds of the three minutes segment were presented. The annotators had the option in each window to generate an apnea annotation by indicating its start and end on the CI.

Sepsis dataset: For this dataset the optimized algorithm was used to annotate CAs from CASEs. We developed a new annotation software using Matlab working as a user interface to resemble the representation returned by a patient monitor and to allow for the identification of CAs. All the characteristics of this software interface are presented in **Figure 5.2**. Two clinical experts provided annotations for apneic events with a central origin using this annotation software. Information extracted from multiple signals (e.g., CI, HR, SpO_2) was delivered to the clinical experts to provide accurate annotations. Different annotation-options were discussed with the clinical experts and the following labels were created: 'central apnea' (a trace of flat CI signal following previously regular fluctuating CI), 'rejection' (flat CI trace, however mixed with large CI fluctuations) and 'error' (missing or corrupt signal).

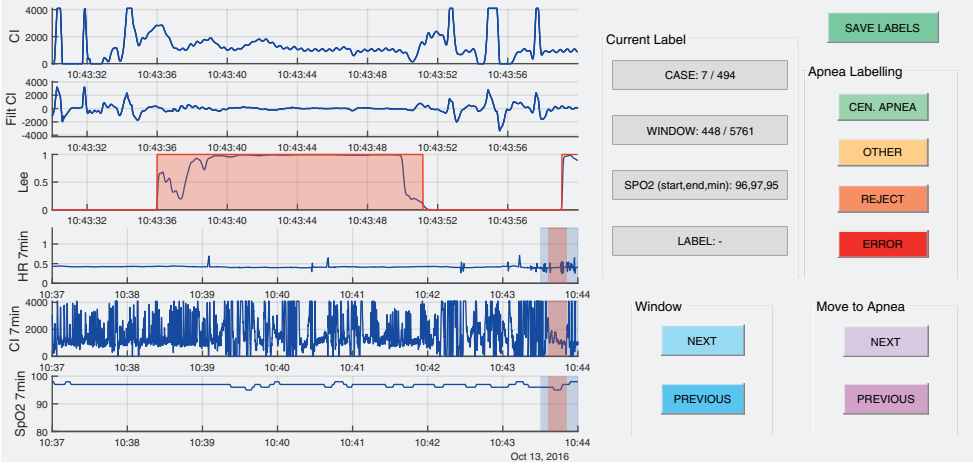


Figure 5.2: Representation of the software annotation used to provide clinical experts' annotations for sepsis dataset based on central apnea suspected events (CASEs) provided by Lee's algorithm. This software includes 30 seconds representations for CI, filtered respiration, apnea probability computed by using Lee's algorithm. RR-intervals, SpO_2 and also CI are represented using a 7-minute window. The starting and ending SpO_2 values registered for each annotation from Lee's algorithm as well as the minimum recorded value found within the annotation are indicated in a grey string.

The current CASE provided by Lee's algorithm is represented in red and is superimposed to all the 7-minute representations. The light blue area is instead used to indicate the position of the 30 seconds windows in the 7-minute representations. The different annotations options are presented on the far right. These are 'central apnea', 'reject' and 'error'. Whenever a choice is made, this is indicated after the string 'label'.

Information about the number of the current CASE as well as the time reference to the whole 48 hours for the selected patients are indicated in gray. 'Save labels' allow the annotator to save all previous annotations.

5.2.5. Data analysis and statistical analysis

Training dataset: All human annotations longer than 10 seconds were identified. Matching apnea annotations were then defined in case the annotations by the different clinical experts presented an overlap for more than 1 second (60 samples). To determine the optimal settings for Lee's algorithm, different cutoffs for the envelope function were evaluated to maximize the level of agreement with matching apnea annotations. Furthermore, the two clinical experts who annotated sepsis dataset were asked to annotate CAs among all matching apnea annotations missed by the algorithm with different cutoffs in the sepsis dataset. The optimal cutoff was defined after clinical consensus considering a trade-off among five parameters: the median counts of CASEs per patient, median recall, median precision, median F1-score (harmonic mean of recall and precision) as well as the total count (sum considering all patients) of missed matching CAs.

Sepsis dataset: The total count of CASEs, matching CAs agreed by both annotators and their median counts per patient were investigated. The count of CASEs, CAs annotated by either annotators (either CAs) and matching CAs was performed using 2-hour windows and a step length of 1 hour. This solution was used to avoid considerations for single hours, since consecutive hours often presented high differences in the count of CASEs.

Then we investigated if the percentage of agreement (ratio between matching CAs and CASEs computed per patient using the same moving window) was influenced by the time-distance from sepsis onset by computing its median value over the different hours. A Cochran-Mantel-Haenszel test was used to evaluate differences between the count of matching CAs found in hours close to sepsis onset and the count found in hours at further time-distance. Differently from a Chi-Square test, this test allows for an analysis of these counts stratified per patient albeit it assumes that the difference in terms of odds ratio is the same across all patients [20].

Data analyses were performed using Matlab for both datasets whereas statistical analyses were performed using R (R Core Team, Vienna, Austria). A p-value < 0.05 was considered significant.

5.3. Results

Training dataset: A total of 259 matching apnea annotations (i.e., annotations of the two different annotators matching by at least one second) were found (median count of matching apnea annotations per patient 21). When running Lee's algorithm on this

dataset, the median count of CASEs was much higher, depending on the cutoff for the envelope function.

Since five parameters related to different cutoffs were analyzed, these will be indicated following the order presented in the method section, for an easier comprehension: median count of CASEs per patient, median recall, median precision, median F1-score and total count of missed matching CAs. First, for the cutoff 0.0025, value suggested by Lee et al. [12] and represented with an orange vertical line in **Figure 5.3**, a median count of 147 CASEs per patient was found (238/259 identified matching apnea annotations), together with a median recall of 94.5%, a median precision of 9.3%, a median F1-score of 16.9% and 3 missed matching CAs in total for all patients. For the cutoff 0.0030, represented with a red vertical line in **Figure 5.3**, the following parameters were extracted: a median count of 144 CASEs per patient (240/259 identified matching apnea annotations), a median recall of 95.5%, a median precision of 9.6%, a median F1-score of 17.4% and 2 missed matching CAs in total. Despite this cutoff showed the maximum median recall, being able to detect most of matching apnea annotations, it also showed a very low precision, structurally detecting a high number of CASEs that were not marked as apneas by the annotators.

To reduce false apnea detection, we aimed for a true to false ratio for CASEs equal to 1:5 (precision = 0.2) and identified 0.0005 as optimal cutoff. For this cutoff, represented with a green vertical line in **Figure 5.3**, the following parameters were extracted: a median count of 75 CASEs per patient (222/259 identified matching apnea annotations), a median recall of 90.5%, a median precision of 19.7%, a median F1-score of 30.8% and 4 missed matching CAs in total. Considering the significant reduction in the count of CASEs returned by the algorithm (improved precision) at the cost of losing just 1 additional matching CA, this cutoff was found to be suitable to annotate large datasets. Despite the median F1-score computed with a cutoff equal to 0.0005 showed a lower value compared to its maximum, as this metric penalizes precision and recall disagreeing with each other too much [21], [22], this value was found to be almost doubled compared to the F1-score computed using the original cutoff, proving to be a suitable choice for the use with sepsis dataset.

Sepsis dataset: A total of 4773 CASEs were detected by the optimized algorithm (median count per patient 537). A total of 772 matching CAs (i.e., CAs agreed on by both annotators) were detected (median count per patient 47).

Time dependencies for the annotations were evaluated by considering the median count of CASEs, either CAs and matching CAs per 2-hour windows, as shown in **Figure 5.4**. A median count of 15 CASEs per patient per 2-hour windows was found, with higher

counts located in proximity to sepsis onset (last 6 hours). The median count of either CAs and matching CAs also showed higher values closer to sepsis.

Percentage of agreement (ratio between matching CAs and CASEs computed per patient) showed higher values in the 6 hours close to sepsis onset as well as very low values in the preceding hours (10-7 hours before sepsis onset). The counts of matching CAs for the last 6 hours prior to sepsis onset were found to be statistically higher than all previous hours by using the Cochran-Mantel-Haenszel test (p-value < 0.05).

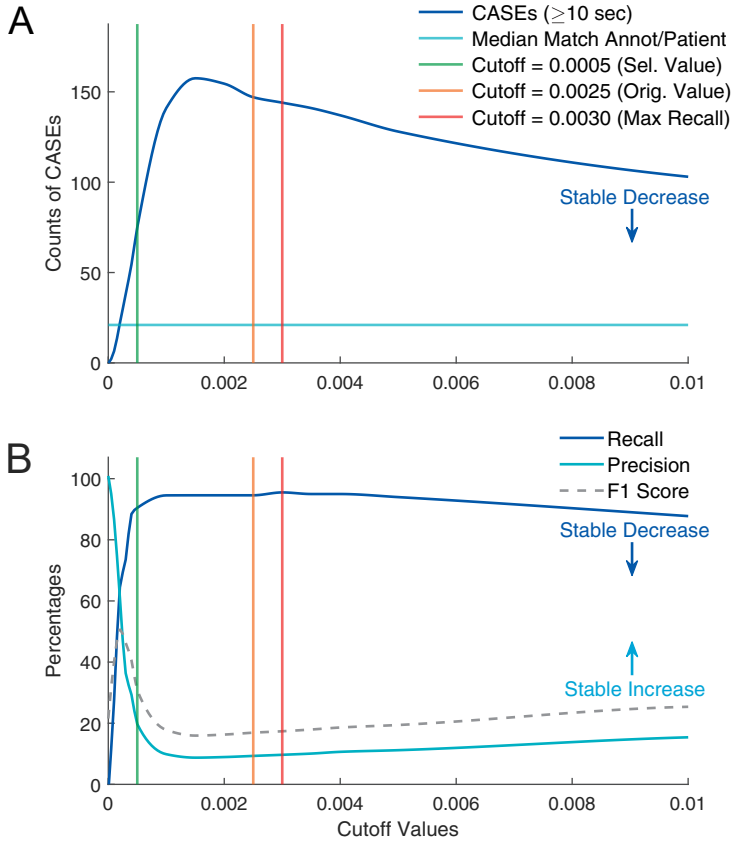


Figure 5.3: (A) Median count of central apnea-suspected events (CASEs) and (B) median precision, recall and F1-score per patient provided by comparing CASEs with matching apnea annotations (provided by clinical experts) from the training dataset. Both figures are generated considering different cutoffs for the envelope function in Lee's algorithm. The optimal cutoff (0.0005) selected for the analysis in sepsis dataset is indicated with a green line. The original choice selected by Lee indicated is with an orange line whereas the cutoff with the highest recall is indicated is with a red line.

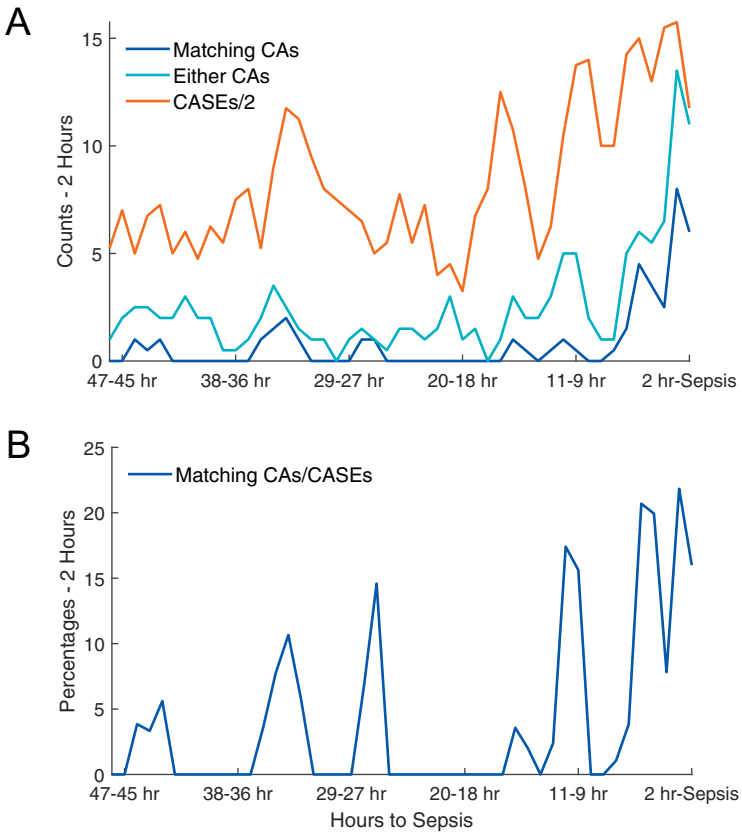


Figure 5.4: (A) Median count of central apnea-suspected events (CASEs) (divided by 2, to allow for scale comparison with the following measures) provided by Lee's algorithm, central apneas (CAs) annotated by either annotators (either CAs) and agreed on (matching CAs) and (B) median percentage agreement computed per patient as ratio between matching CAs and CASEs. All measures were computed using 2-hour windows and a step length of 1 hour.

5.4. Discussion

In this study we optimized an automatic algorithm to facilitate human annotations for CAs in a dataset of premature infants and used it to detect CAs preceding LOS and to evaluate agreement for CAs before LOS. This study proposes a method to tune Lee's algorithm for CA detection for the use in diverse and large datasets of premature infants.

Algorithm optimization was possible by establishing an optimal cutoff for the envelope function (0.0005). This choice provided only a slight loss of recall (90.5% vs. 94.5%, median values per patient) but resulted in an improved precision (19.7% vs 9.3%) and F1-score (30.8% vs. 16.9%). The optimized algorithm returned therefore less than half CASEs per

patient for further clinical annotations compared to when the original cutoff was used, significantly reducing the time needed by clinical experts to annotate a dataset. Since availability of clinical experts is always limited, we considered the increase of precision a valuable improvement for the algorithm. This result was obtained at the price of missing just one additional matching CA in the whole training dataset (4 vs 3).

A software annotation tool was created in close contact with clinicians to allow for the detection of CAs among all CASEs. The solution proposed in this work considers both the need to present information in a similar way to the one commonly shown by patient monitors and to include relevant information for the detection of CAs.

The influence of the time to sepsis on CASEs and CAs was then evaluated. The optimized algorithm returned a higher count of CASEs per 2-hour windows in the last 6 hours prior to sepsis onset. The count of matching CAs in the last 6 hours was found to be statistically higher when compared to all previous hours (p-value < 0.05). This difference was also reflected by the percentage of agreement, a result which could indicate that apnea characteristics may be more pronounced close to sepsis. Previous works showed that sepsis is preceded by an increase in the number of apneic periods [7]–[9]. A higher agreement for CAs in these hours can therefore be motivated by the association between sepsis and apneas. Interestingly, the effect of sepsis is thus already visible 6 hours before its onset: future studies may investigate if the presence of a higher number of CAs may be relevant for the prediction of sepsis onset while using different machine learning algorithms.

This study has limitations. First, only the cutoff parameter of the envelope function of Lee's algorithm was optimized. The possibility of using different functions from the two-parameter sigmoid function to compute apnea probability could be further examined. Second, in this work we only focused on CA annotations but both annotators clearly reported that 'rejected' CASEs often would have been categorized as mixed or obstructive apneas. An evaluation of the precision and recall for this type of events could provide additional value to the optimized algorithm.

5.5. Conclusion

Annotation of CAs in a large dataset of premature infants is feasible by using the optimized algorithm, which returns CASEs for further human annotations. This optimized algorithm proved to be reliable for annotating CAs (very few CAs missed) while also reducing the time needed by clinical experts to annotate a dataset. The identification of high counts of CAs seems a promising tool to provide additional useful information to help clinical experts predict the onset of sepsis in premature infants.

References

[1] G. Schmalisch, "Neonatal monitoring," *Capnography, Second Ed.*, no. 2012, pp. 80–95, 2011.

[2] N. N. Finer, R. Higgins, J. Kattwinkel, and R. J. Martin, "Summary proceedings from the apnea-of-prematurity group," *Pediatrics*, vol. 117, no. 3, pp. S47–S51, 2006.

[3] E. C. Eichenwald and Committee on Fetus and Newborn, "Apnea of Prematurity," *Pediatrics*, vol. 137, no. 1, 2016.

[4] J. Zhao, F. Gonzalez, and D. Mu, "Apnea of prematurity: From cause to treatment," *Eur. J. Pediatr.*, vol. 170, no. 9, pp. 1097–1105, 2011.

[5] J. M. Di Fiore, R. J. Martin, and E. B. Gauda, "Apnea of prematurity - Perfect storm," *Respir. Physiol. Neurobiol.*, vol. 189, no. 2, pp. 213–222, 2013.

[6] A. Janvier, M. Khairy, A. Kokkotis, C. Cormier, D. Messmer, and K. J. Barrington, "Apnea is associated with neurodevelopmental impairment in very low birth weight infants," *J. Perinatol.*, vol. 24, no. 12, pp. 763–768, 2004.

[7] K. Fairchild *et al.*, "Clinical associations of immature breathing in preterm infants: Part 1-central apnea," *Pediatr. Res.*, vol. 80, no. 1, pp. 21–27, 2016.

[8] L. Cabrera-Quiros *et al.*, "Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning," *Crit. Care Explor.*, vol. 3, no. 1, p. e0302, 2021.

[9] M. H. Tsai *et al.*, "Recurrent late-onset sepsis in the neonatal intensive care unit: Incidence, clinical characteristics and risk factors," *Clin. Microbiol. Infect.*, vol. 20, no. 11, pp. O928–O935, 2014.

[10] B. Stoll *et al.*, "Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network," *Pediatrics*, vol. 110, no. (2 Pt 1), pp. 285–91, 2002.

[11] R. J. Martin, J. M. Abu-Shaweesh, and T. M. Baird, "Apnoea of prematurity," *Paediatr. Respir. Rev.*, vol. 5, no. SUPPL. A, 2004.

[12] H. Lee *et al.*, "A new algorithm for detecting central apnea in neonates," *Physiol. Meas.*, vol. 33, no. 1, pp. 1–17, 2012.

[13] S. Picone, R. Aufieri, and P. Paolillo, "Apnea of prematurity: challenges and solutions," *Res. Reports Neonatol.*, p. 101, 2014.

[14] M. A. Mohr *et al.*, "Very long apnea events in preterm infants," *J. Appl. Physiol.*, vol. 118, no. 5, pp. 558–568, 2015.

[15] S. Y. Belal, A. J. Emmerson, and P. C. W. Beatty, "Automatic detection of apnoea of prematurity," *Physiol. Meas.*, vol. 32, no. 5, pp. 523–542, 2011.

[16] J. R. Williamson, D. W. Bliss, D. W. Browne, P. Indic, E. Bloch-Salisbury, and D. Paydarfar, "Using physiological signals to predict apnea in preterm infants," *Conf. Rec. - Asilomar Conf. Signals, Syst. Comput.*, pp. 1098–1102, 2011.

[17] B. D. Vergales *et al.*, "Accurate automated apnea analysis in preterm infants," *Am. J. Perinatol.*, vol. 31, no. 2, pp. 157–162, 2014.

[18] K. D. Fairchild, "Predictive monitoring for early detection of sepsis in neonatal ICU patients," *Curr. Opin. Pediatr.*, vol. 25, no. 2, pp. 172–9, 2013.

[19] N. Kumar, G. Akangire, B. Sullivan, K. Fairchild, and V. Sampath, "Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront," *Pediatr. Res.*, vol. 87, no. 2, pp. 210–220, 2020.

[20] A. Agresti, *An introduction to categorical data analysis*, 2nd ed. John Wiley & Sons, 2007.

[21] Y. Sasaki, "The truth of the F-measure," no. January 2007, pp. 1–6, 2015.

[22] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Mach. Learn.*, vol. 110, no. 3, pp. 451–456, 2021.

CHAPTER 6

Insights into mechanisms underlying central apnea

Reproduced after: The effect of apnea length on vital parameters in apnea of prematurity – Hybrid observations from clinical data and simulation in a mathematical model, Early Human Development, 2022, pp. 1-8.

Abstract

Apnea of prematurity (AOP) is a critical condition for preterm infants which can lead to several adverse outcomes. Despite its relevance, mechanisms underlying AOP are still unclear. In this work we aimed at improving the understanding of AOP and its physiologic responses by analyzing and comparing characteristics of real infant data and model-based simulations of AOP.

We implemented an existing algorithm to extract apnea events originating from the central nervous system from a population of 26 premature infants (1248 hours of data in total) and investigated oxygen saturation (SpO_2) and heart rate (HR) of the infants around these events. We then extended a previously developed cardiovascular model to include the lung mechanics and gas exchange.

After simulating the steady state of a preterm infant, which successfully replicated results described in previous literature studies, the extended model was used to simulate apneas with different lengths caused by a stop in respiratory muscles. Apneas identified by the algorithm and simulated by the model showed several similarities, including a far deeper decrease in SpO_2, with the minimum reached later in time, in case of longer apneas. Results also showed some differences, either due to how measures are performed in clinical practice in our neonatal intensive care unit (e.g., delayed detection of decline in SpO_2 after apnea onset due to signal averaging) or to the limited number of very long apneas (≥ 80 s) identified in our dataset.

6.1. Introduction

Apnea of prematurity (AOP) is very common in preterm infants (gestational age < 37 weeks) and is almost universal in low-birth weight infants below 1000 g [1]. It has been associated with later morbidity and other adverse outcomes [2], [3]. Although definitions for apnea events vary, it is generally defined as a cessation of breathing lasting more than 20 s, or a cessation of breathing lasting more than 10 s accompanied by bradycardia (heart rate, HR < 100 bpm), desaturation (oxygen saturation, SpO_2 < 80%) or both [1]. In some apnea events, bradycardia and desaturation occur almost immediately, while in other events these responses are delayed or even absent [4]. The length of the apnea, period in which no respiration occurs, influences the depth of the bradycardia and desaturation that are associated with the apnea. In addition, increased apnea lengths have been associated with a slower decrease in SpO_2, possibly due to higher initial SpO_2 values [5]. An initial low value of arterial SpO_2 is thought to cause the rapid fall in SpO_2 in recurrent apneas, when a second apnea follows closely on a first apnea [6]. However, in some cases a decrease in HR occurs instantaneously with apnea before a fall in SpO_2 has occurred or even in the absence of a fall in SpO_2, which suggests a central mechanism for bradycardia [7].

Mechanisms underlying bradycardia during apnea are still unclear. AOP is believed to be a manifestation of physiologic immaturity of the respiratory control [2]. The respiratory center regulates a stable breathing pattern that adapts accurately to changes in metabolic need [8]. This center receives excitatory and inhibitory inputs from 1) higher brain centers responding to, for example, emotional or volitional stimuli, 2) mechanoreceptors in the respiratory tract avoiding overinflation of the lungs, 3) peripheral chemoreceptors in the carotid and aortic body sustaining O_2-, CO_2- and pH-levels, and 4) central chemoreceptors on the ventral medullary surface sustaining CO_2- and pH-levels [9]. Immaturity is thought to cause impaired breathing responses to hypercapnia and hypoxia, respectively, contributing to the occurrence and severity of AOP.

In this study, we aimed at improving the understanding of apnea and its physiologic responses in premature infants. First, we implemented an automatic detection of cessations of breathing, based on a previously published algorithm [10]. We analyzed apnea events by studying clinical patient monitoring data from preterm infants admitted to our neonatal intensive care unit. In the clinical data we investigated the effect of cessation of chest respiration motion on the SpO_2 and HR of the infant. Second, we extended a previously developed cardiovascular model [11] with a lung mechanics model and gas exchange model in which we simulated periods of varying lengths of cessation of breathing and investigated the effect on SpO_2 and HR. We hypothesized that

results from physiological data and simulations by means of mathematical modeling can enhance understanding of AOP and its etiology by confirming dependencies between presumably related variables.

6.2. Methods

6.2.1. Study population

All patients included in this study consisted of 26 very low-birth-weight infants (birth weight < 1500 g) characterized by a gestational age < 30 weeks. For each patient, 48 h of data were extracted and used in this study. Patient demographics are presented in **Table 6.1**.

All selected patients had their ECG, chest impedance (CI) and SpO_2 collected using the Philips IntelliVue MX800 patients' monitors (Boeblingen, Germany), according to clinical standard. ECG and CI were measured using three ECG leads. Sample frequency for the ECG was 250 Hz, whereas for the CI was 62.5 Hz. SpO_2 was captured from by PPG at 1 Hz. The data was stored in a data warehouse (Data warehouse connect, release B.01, Philips Medical Systems, Andover, USA) and pseudonymised before storage for further analysis.

As this study had a retrospective non-invasive nature, a waiver was provided by the medical ethical committee in accordance with the Dutch law on medical research with humans (WMO).

6.2.2. Data analysis

Patients' ECG and CI signals were processed using a central apnea detection algorithm described by Lee et al. [10]. This algorithm allows to filter out all the contributions given by cardiac artifact from the original respiration to create a filtered respiration signal. Consequently, this algorithm returns a central apnea probability function from which all events characterized by a cessation of breathing longer than 10 s in the CI can be identified as central apnea-suspected event (CASEs). The length of each CASE corresponds to that of a cessation of breathing present in the original CI signal. A more extensive description on the original algorithm is found in [10]. For this work, as an additional step to the Lee's algorithm, all CASEs within 10 s distance were combined, in line with clinical practice, where cessations of breathing close to each other are evaluated as one event knowing that the decrease in SpO_2 can be amplified by different consecutive apneas. All CASEs followed by a decrease in SpO_2 within the start of each CASE and the two minutes following its end were identified. All instances where $SpO_2 \leq$ 85% were defined as so-called CASE-Ds (with D indicating desaturation). Definitions for

desaturation differ between $SpO_2 \leq 80\%$ and $SpO_2 \leq 85\%$, but since the latter is already associated with negative outcomes [4] we recently implemented a new type of alarm for a decrease in SpO_2 lower than 85% as part of alarm optimization process in our neonatal intensive care unit [12]. As such we classified all instances of $SpO_2 \leq 85\%$ as desaturation.

CASEs and CASE-Ds were then divided and grouped according to the length of their cessation of breathing L. The following lengths L were considered: 10 s ≤ L < 20 s, 20 s ≤ L < 30 s, 30 s ≤ L < 40 s, 40 s ≤ L < 50 s, 50 s ≤ L < 60 s, 60 s ≤ L < 80 s, L ≥ 80 s. The subdivision was selected based on the suggestions for grouping apnea-bradycardia-desaturations proposed by Mohr et al. [5] and expanded even further to elaborate the analysis of events based on their length. As a final revision for CASEs and CASE-Ds used in our work, all CASEs that were preceded by another CASE longer than 10 s in the previous 120 s were excluded from the analysis. This characteristic was introduced to avoid the inclusion of modifications in patient signals due to the contributions of previous CASEs or CASE-Ds. Removal of CASEs and CASE-Ds preceded by CASEs shorter than 10 s was not considered necessary as these events do not possess the characteristics of an apnea as defined in [1]. We instead calculated the percentage of CASEs and CASE-Ds preceded in the previous 10 s by at least one short CASE (L < 10 s) in order to estimate the influence of this effect. For each CASE-D, HR (derived from the ECG) and SpO_2 were extracted. Mean values for these two signals were then computed by considering all CASE-Ds included in each group.

Table 6.1: Characteristics of patients included in the analysis.

Count of patients	26
Gestational age (weeks)	27.7 ± 1.0
Postnatal days	9.9 ± 7.4
Birth weight (g)	1039.0 ± 194.7
Gender (female)	46.1%
Hours per patient considered	48
Total hours	1248

6.2.3. Cardiorespiratory model

The cardiorespiratory mathematical model used in this work is an extension of previously implemented models [11]. In particular, one of these recent models was developed to simulate preterm infant physiology and consisted of a cardiovascular component with a baroreceptor model [11]. In this work we extended this model with a lung mechanics component and a gas exchange model to include the respiratory system. An overview of how these components are structured is shown in **Figure 6.1A**.

The cardiovascular model presented in [11] is a lumped parameter model of the neonatal systemic and pulmonary circulation, adapted from Sa Couto et al. [13]. Cardiac contractile cavities and vascular compartments are described as nodes with the ability to store blood. In the contractile cavities the elastance E is varied during the heart cycle to simulate the contraction of the heart, causing the stored blood to flow out into systemic and pulmonary circulation. Vascular compartments are characterized by a constant compliance C and an unstressed volume V_0. These nodes are connected with segments that described flow between nodes based on pressure difference Δp over the segment and a resistance R of that segment, dependent on the size of the vessel. Regulation by a baroreceptor is modeled according to Wesseling et al. [14] using the parameters provided by Ursino et al. [15], [16] and was adapted for a premature infant [11]. The baroreceptor model was introduced to describe the model response to a deviation in arterial blood pressure from reference value for pressure.

A model for the lung mechanics was introduced based on the work from Albanese et al. [17], but simplified to one dead space compartment, combining the larynx, trachea and bronchi (**Figure 6.1B**) to turn it into a 3 compartments model with lungs, dead space and chest wall. Muscle pressure is modeled as a periodically changing pressure. This generated pressure changes inside the lungs, driving the air flows in and out, thus creating changes in the volumes. A more in-depth description of this model is presented in **Appendix 6.1**.

Gas exchange was modeled within the lung compartment as one alveolus (**Figure 6.1C**). The amount of oxygen stored in a node is determined by oxygen concentration in the blood bound to hemoglobin and the difference in oxygen pressure between alveoli and pulmonary microcirculation. Oxygen saturation SpO_2 is then calculated from the partial oxygen pressure pO_2 according to the Hill equation [18]. Further details about this model are presented in **Appendix 6.2**. The amount of oxygen VO_2 in a node changed due to convective transport Q_C, diffusive transport Q_D, and metabolic rate Q_m in the systemic arterial micro-circulation. Oxygen uptake occurred in the pulmonary micro-circulation and depended on capillary gas exchange parameters.

6.2.4. Model simulations

A Euler forward integration scheme was used to solve the model numerically. The model was solved for every time step Δt. Changes in air and blood volume, pressure and flow were calculated by taking the volume at the current time point t_n as a starting point.

All cardiovascular and baroreceptor parameters were estimated to simulate a one week old preterm infant with a gestational age of 28 weeks and birth weight of 1000 g, using scaled parameters according to Jennekens et al. [11], who scaled values for a full term infant from Sa Couto et al. [13]. Lung mechanics parameters were used from Albanese

et al. [17] and scaled down similar to Jennekens' method [11]. Parameters for the oxygen model were based on literature [13], [17]–[20]. The cardiac output was estimated as 175 ml/min and the total blood volume of the infant at 110 ml [11]. Parameters for both the lung mechanics and gas exchange models are shown in **Appendix 6.3**.

Before simulating apnea with our model, we investigated whether the steady state of the model simulated a preterm infant adequately. Therefore, we modeled a preterm infant until steady state had been reached. Hemodynamic outputs of the model were compared with reference values provided in [21], [22]. Particular attention was also dedicated to the parameters for the respiratory model as well as the oxygen levels as these are especially relevant for the modeling of apnea: these parameters were compared with the reference values from [23]–[25].

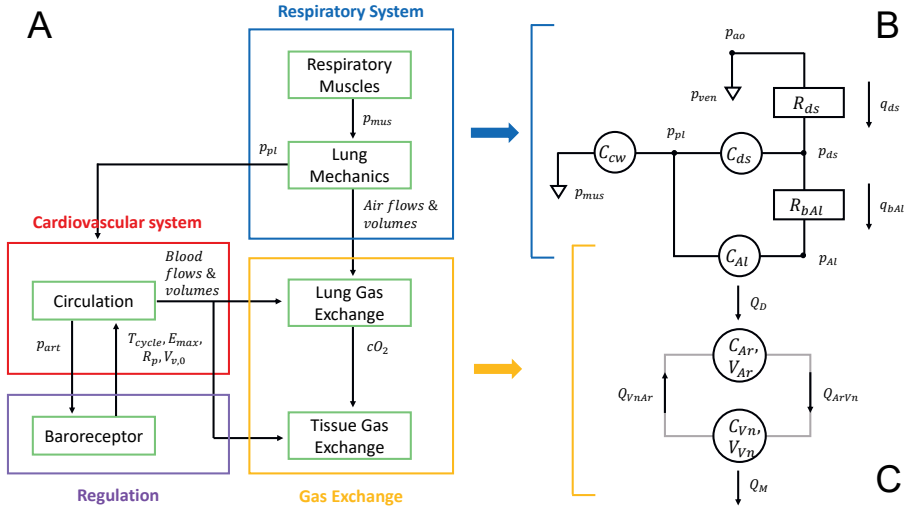


Figure 6.1: Schematic of the cardiorespiratory model used in this work:

(A) Representation of the different models included in the cardiorespiratory model including their interaction. The cardiovascular system consists of the circulation which is regulated by the baroreceptor [13], [14]. Blood flows and volumes, generated by the cardiovascular module, influence diffusion and metabolic rate. Diffusion rate is also dependent on oxygen flow in the lungs, caused by movement of the respiratory muscles [17]. Varying pleural pressures due to respiration affect pressures in intra-thoracic components;

(B) Lung mechanics model: simplified from Albanese et al. [17] into one compartment. Pressures p are indicated at the nodes. Segments represent resistances R and nodes compliances C. Flows over the resistances are visualized with the arrows;

(C) Gas exchange model: subscripts: ao – airway opening, ds – dead space, bAl – bronchi to alveoli, Al – alveoli, pl – pleura, cw – chest wall, mus – respiratory muscles, ven – ventilator. Abbreviations: cO2 – oxygen concentration, Emax – maximum elastance, part – arterial pressure, pmus – muscle pressure, ppl – pleural pressure, Rp – peripheral resistance, Tcycle – cardiac cycle time, Vv,0 – unstressed venous volume.

Apnea was simulated by adjusting lung mechanics parameters to simulate a cessation of breathing caused by a stop in respiratory muscles (i.e., respiration rate, RR = 0), similarly to what happens in premature infants for apneas with a central origin. The model was run for multiple apnea lengths to study the effect of apnea length on vital signs, and the outcomes of the model were studied and compared with CASE-Ds from real infants' data. Cessations of breathing with different lengths L were included in this work to mimic the apnea periods from the data study. We therefore chose to model apnea lengths of 10, 20, 30, 40, 50, 60 and 80 s.

Apnea was also simulated after varying the value of different parameters: these included the initial respiration rate (RR), metabolic rate (Q_m) and alveolar compliance (C_{Al}). In literature, an increase in oxygen levels preceding very long apneas caused by an increase in RR was described by Mohr et al. [5]. They hypothesized that this increase in SpO_2 caused a decrease in the slope of SpO_2. To verify this finding in our model we simulated increased SpO_2 values before apnea by increasing the RR at start. Reference RR, set equal to 49 bpm, was multiplied by factors 0.9, 1, 1.1 and 1.2. These factors were estimated so that SpO_2 values laid in the range reported in [5], as no average RR were reported. Outcomes obtained with different RR were then studied and compared with each other. Initial Q_m was set equal to 0.12 mlO_2/s, as estimated by previous works considering a population of preterm infants [26], [27]. Different apneas were then generated using the model by multiplying this value by factors 0.5, 0.8, 1 and 1.2, selected to consider a wide number of possibilities for this parameter. Different C_{Al} were evaluated due to the influence of alveolar stiffness on the tidal volume of the premature infants, a characteristic which could significantly affect the simulated apneas. C_{Al} was initially set equal to 0.93 ml/cmH_20 following [17] and multiplied by factors 0.5, 0.8, 1 and 1.5 to estimate its contribution on the simulated apneas.

6.3. Results

The total count of CASEs and CASE-Ds is presented in **Table 6.2**. The count of events for each group decreased when longer lengths L were considered. This table also reports the median count and percentiles [25, 75] of CASE-Ds computed per patient as well as the median ratio between the count of CASE-Ds and CASEs. 64.5% of CASEs and 59.7% of CASE-Ds included in this work were preceded by at least one short CASE in the 10 s preceding their onset.

A mean representation of HR and SpO_2 for CASE-Ds considering the different L is presented in **Figure 6.2**. All mean curves for both HR and SpO_2 for CASE-Ds with L < 80 s (**Figure 6.2A-B**) shared the property that their minimum value is reached later in

time when longer CASE-Ds are considered. For each length L, the minimum HR value was always occurring earlier in time compared to the minimum SpO_2 value. CASE-Ds with L ≥ 80 s showed different patterns from what was shown by shorter CASE-Ds. Curves for mean HR and SpO_2 for CASE-Ds with L ≥ 80 s (**Figure 6.2C-D**) presented several oscillations which were not present in shorter CASE-Ds, possibly indicating the contribution of multiple shorter CASEs. An additional analysis computed separately for each CASE-Ds with L ≥ 80 s showed brief moments of recovery from cessations of breathing for different patients. These interruptions between long periods without breathing related motion in the CI signal explain why there are some prominent peaks and troughs in the mean curves.

The cardiorespiratory and oxygen models were used to simulate a stable preterm infant, for which results are presented in **Figure 6.3**. The outcomes of this new combined model showed that the oxygen pressures in the arteries and veins remained constant at approximately 60 and 32 mmHg, respectively (**Figure 6.3A**). SpO_2 levels in lung microcirculation and pulmonary vein reached 95%, since parameters were adjusted to match this value (**Figure 6.3B**). The volume of the alveoli fluctuated with a tidal volume of 2.3 ml, while dead space remained mostly constant (**Figure 6.3C**).

Apnea was simulated by generating cessation of respiration motion at t = 0 s. Results for these experiments, obtained with different apnea lengths are presented in **Figure 6.4A**. During apnea, arterial SpO_2 decreased, with a total decrease in SpO_2 of 5%, 12%, 17%, 22%, 27%, 32% and 43% for apnea lengths L of 10, 20, 30, 40, 50, 60 and 80 s, respectively. During apnea, the slope of SpO_2 was independent of length L. However, the rate of recovery of the arterial SpO_2 increased with increasing L. Simulations while varying the RR, Q_m and C_{AI} are shown in **Figure 6.4B-D**. Higher respiration rate (RR) caused an increase in the initial SpO_2 which was respectively found to be equal to 92%, 94%, 95% and 96%, as shown in **Figure 6.4B**. Higher initial RR resulted in a slower decrease in the initial phase after the apnea and to a higher rate of recovery of the arterial SpO_2. Higher metabolic rate (Q_m) resulted in a progressively steeper decrease in SpO_2 and led to progressively lower minimum values which were equal to 89%, 74%, 63% and 51%, respectively. SpO_2 recovery was also affected by this parameter. The effect of different C_{AI} on the generated apneas was very limited: higher C_{AI} mildly affected the initial SpO_2 but had no influence neither on SpO_2 decay nor on its recovery after the minimum was reached.

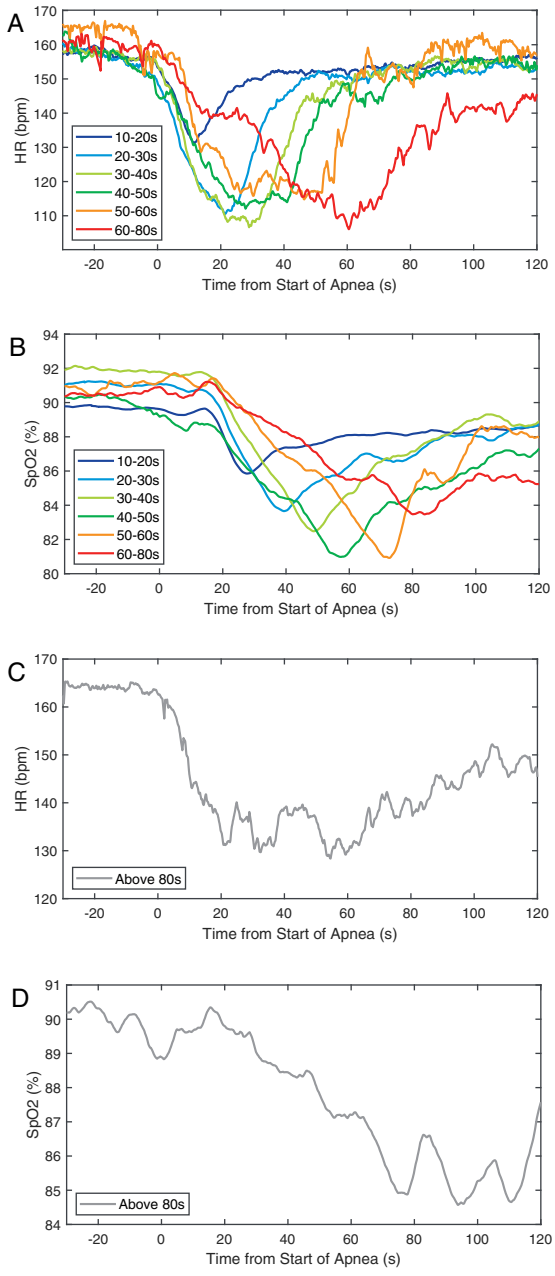


Figure 6.2: Mean representation of heart rate (HR) and oxygen saturation (SpO_2) curves for central apnea-suspected annotations followed by a decrease in SpO_2 ≤ 85%(CASE-Ds) considering the different lengths of the cessation of breathing L. HR and SpO_2 curves for CASE-Ds with 10 s ≤ L < 20 s, 20 s ≤ L < 30 s, 30 s ≤ L < 40 s, 40 s ≤ L < 50 s, 50 s ≤ L < 60 s, 60 s ≤ L < 80 s are presented in (A) and (B). HR and SpO_2 curves for CASE-Ds with L ≥ 80 s are presented in (C) and (D). All patients included in the dataset are included in these representations.

Table 6.2: Count of central apnea-suspected annotations (CASEs) and central apnea suspected annotations followed by a decrease in oxygen saturation (SpO_2) ≤ 85% (CASE-Ds), median and percentiles [25,75] of CASE-Ds and median ratio CASE-Ds/CASEs computed per patient. All values are presented for different lengths for the cessations of breathing L.

Measure	10s≤L <20s	20s≤L <30s	30s≤L <40s	40s≤L <50s	50s≤L <60s	60s≤L <80s	L≥80s	All
# CASEs	3140	655	404	202	106	127	122	4756
# CASE-Ds (SpO_2 ≤ 85)	646	190	120	60	31	43	28	1118
Median CASE-Ds	13.5	5.5	2.5	1	0	1	1	25.5
Percentiles [25,75] CASE-Ds	[3, 43]	[1, 11]	[1, 7]	[0, 3]	[0, 2]	[0, 3]	[0, 1]	[10, 74]
Median ratio CASE-Ds/ CASEs	0.13	0.26	0.25	0.20	0.24	0.47	0.20	0.17

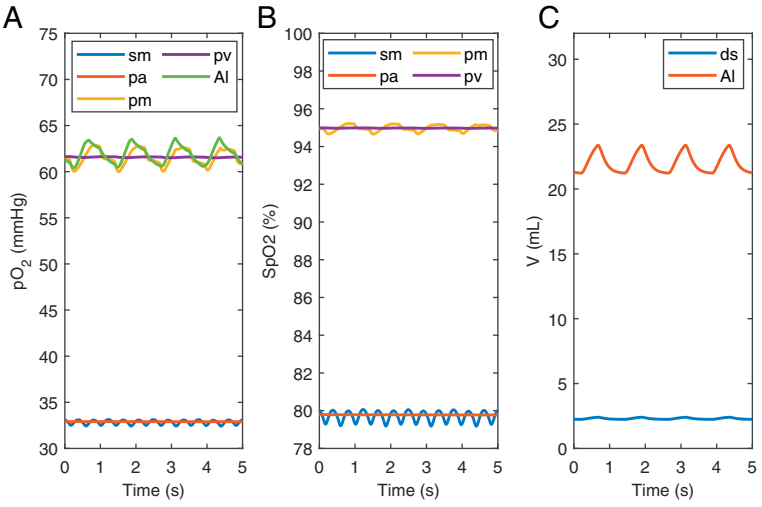


Figure 6.3: (A) Simulated oxygen partial pressure (pO_2), (B) oxygen saturation (SpO_2) and lung volume V (C) as a function of time for a preterm infant according to original model under baroreceptor control. Abbreviations: AI – alveolar, ds – dead space, pa – pulmonary artery, pm – pulmonary micro-circulation, pv – pulmonary vein, sm – systemic micro-circulation.

All apnea simulations also presented a common pattern for the HR: after a slight decrease that follows the start of the apnea, HR increased from 142.2 to 143.0 bpm irrespectively of the apnea lengths. When apnea ended, HR decreased to its original value of 142.2 bpm. This effect is due to the baroreceptor but is not shown by our results obtained with clinical data.

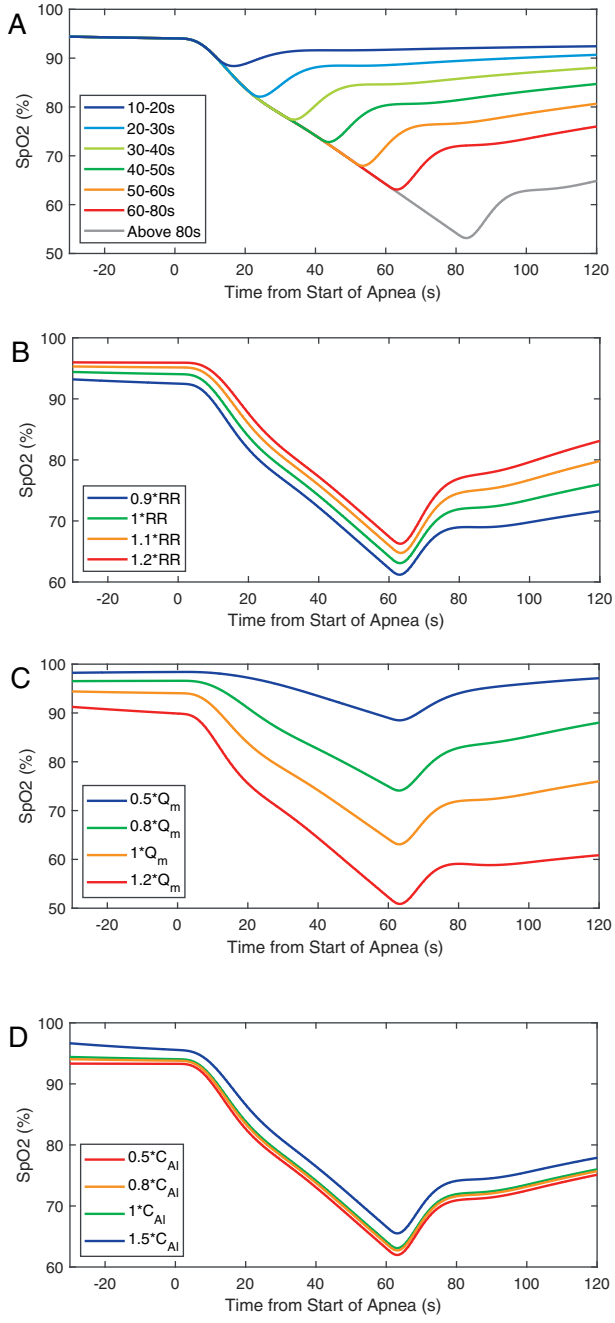


Figure 6.4: Oxygen saturation (SpO_2) curves obtained by simulating apneas with different lengths of the cessation of breathing L (A) and by varying the initial respiration rate (RR) (B), metabolic rate (Q_m) (C) and alveolar compliance (C_{Al}) (D). Variations of the different parameters that were evaluated are reported in each figure. In all figures apnea starts at t = 0 s by setting RR = 0. Reference RR is equal to 49 bpm, reference Q_m is equal to 0.12 mlO_2/s and compliance C_{Al} is equal to 0.93 ml/cmH_20.

6.4. Discussion

The analysis performed with our population of preterm infants showed interesting similarities with the work from Mohr et al. [5].

A first result shared by both works indicates that the observed decline in HR starts most commonly before the decline in SpO_2 [5]. This result is also in accordance with other previous works including the one from Carbone et al. in which it is mentioned that a drop in HR preceded the one in SpO_2 in 57% of long apnea (L ≥ 20 s), in 56% of medium apnea (15 – 19 s), and in 41% of short apnea (10 – 14 s) [4]. Joshi et al. in a study related to time-relationship among cessation of breathing, bradycardia and desaturation alarms also reported that bradycardia has a much higher likelihood of being followed by desaturation (37% vs. 11.5%) than vice versa and indicated a much shorter median time for the transition from one event to the other (0.7 min vs. 18 min) [28]. However, it is difficult to draw any conclusion from these observations because of two time-delays in the measurement of SpO_2. The measuring device has an averaging time set in our practice to 10 s and there is a circulation delay between blood in the lungs and blood measured in the hand or foot, which may be around 6 – 8 s in our infants. This information can be estimated considering that the blood volume in premature infants is around 65 – 80 mL/Kg [29], [30] and the cardiac output is around 0.23 l/min [31], corresponding to a total circulation time (blood volume/cardiac output) of around 20 s. Since the blood volume in veins is about twice the amount present in arteries, the circulation delay can be estimated in the vicinity of 7 s. Thus, the total delay in the two measurements may be around 17 ± 2 s. This is consistent with the observations presented in **Figure 6.2**, which show SpO_2 beginning to fall around 17 s after the cessation of breathing. It follows that we cannot conclude just from these observations whether the drop in HR actually begins before or after the drop in SpO_2.

Differently from what is described in [5], we observed the decline in HR starting slightly before the cessation of breathing. We do not expect this to be a result of timing errors in the monitoring system. We found however that 59.7% of CASE-Ds were preceded by short CASEs in the previous 10 s (indicating the presence of cessations of breathing shorter than 10 s). We acknowledge that short cessations of breathing preceding the actual CASE-Ds could have contributed to a drop in HR in multiple occasions but nonetheless we did not consider the exclusion of CASE-Ds preceded by short CASEs from our study since these short CASEs do not comply with the definition of apnea presented in [1]. The drop in HR happening before the cessation of breathing indicates that this drop in HR could not be caused by the drop in SpO_2 since the decrease in saturation is a direct consequence of the interruption in the breathing motion. The drop in HR preceding the cessation of breathing suggests therefore that other mechanisms

may be involved, such as a central mechanism for bradycardia, or a central mechanism producing both bradycardia and apnea. However, the presence of short CASEs prior to the long apneas studied in this work makes it difficult to draw any definitive conclusion.

Other important similarities with the work of Mohr et al. included minimum SpO_2 values appearing later in time and lower values in case of longer cessations of breathing, for all L < 80 s. In addition, longer decreases in HR were also found in case of longer cessations of breathing. Mohr at al. suggested that the slower decrease in SpO_2 for longer cessations of breathing could be related to the elevated SpO_2 values that they found before very long apneas, a characteristic which was possibly caused by an increase in RR [5]. This hypothesis was supported by a model analysis of arterial oxygen desaturation during apnea in a preterm infant by Sands et al. [6]. However, we cannot confirm elevated SpO_2 values before the start of longer cessations of breathing in our dataset.

By looking at apneas extracted from our population of premature infants we also noticed that the rate of decrease in SpO_2 seemed to slow down as the length of the cessation of breathing increased. We hypothesize that the respiratory centers might be triggered by the rate of decrease in SpO_2, leading to a faster trigger of the respiratory muscles in case of higher rates of decrease, typical of shorter apneas, and thus stopping apnea with a quicker response. Recovery of SpO_2 values found before the start of the cessation of breathing was also found to happen earlier in time in case of shorter CASE-Ds, similarly to what is presented in [5].

In our study, HR and SpO_2 curves for CASE-Ds longer than 80 s showed different patterns compared to the curves from shorter CASE-Ds. We observed for L ≥ 80 sa smaller decrease in SpO_2 than for shorter cessations of breathing. This result can be explained by the fact that these CASE-Ds were often the result of the combination of short CASEs and in the time frame between each CASE the patient had possibly taken a breath to adjust SpO_2 levels. This result was verified with a separate analysis of each CASE-Ds longer than 80 s, which indicated an overlap for the most prominent peaks and troughs in the HR and SpO_2 in both single CASE-Ds and in the mean curves included in **Figure 6.2C-D**. Furthermore, these modification in HR and SpO_2 resulted being associated to brief interruptions in the cessations of breathing, proving that CASE-Ds longer than 80 s often resulted from the combination of short CASEs. Note that this combination did not frequently occur in the other CASE-D groups, as mean representations included in **Figure 6.2A-B** did not include the same fluctuations in HR and SpO_2 as shown in **Figure 6.2C-D**.

Our simulations in the stable premature cardiovascular system showed that, with the addition of the respiratory models, our model was able to simulate the physiology of

a premature infant as values calculated in our model are in line with reference values from literature for both cardiovascular parameters [21], [22], [32], [33] and for oxygen and partial oxygen pressures [23]–[25]. An exception, however, is the venous SpO_2 level which showed higher values than those found in literature [23]. This might be attributed either to the fact that the values found in literature refer to preterm infants with different postnatal ages or to the choices made in the model with respect to parameters, as not all parameters can be easily measured, and they are estimated based on literature studies.

In response to the different simulated apneas, the total decrease in SpO_2 was smaller in shorter apneas, as breathing resumed earlier. Recovery of the initial SpO_2 values was also found to happen faster in case of shorter apneas. This result reflects the findings in the data from our population of preterm infants for CASE-Ds shorter than 80 s, and they are in line with values found in literature [4], [34].

During apnea simulation, arterial SpO_2 showed a biphasic response in our model, a result not easily recognizable in our data. Results from the model showed that initially SpO_2 decreased rapidly. This steep decrease was then followed by a slower decrease. This biphasic response to apnea is not widely discussed in literature, however, it was observed in some studies [35], [36]. Later, this phenomenon was investigated and confirmed in other studies [6], [37]. In the model, systemic arterial oxygen pressure closely followed alveolar oxygen pressure. Under normal conditions, systemic venous oxygen pressure was about 30 mmHg lower. During early apnea, this difference in oxygen pressure led to a large diffusive rate that quickly drained oxygen from the alveoli causing both alveolar and arterial pressure to drop. When alveolar and arterial oxygen pressure approached venous oxygen pressure, essentially alveoli and the blood could be considered to behave as one single oxygen storage compartment, from which oxygen was consumed at a constant metabolic rate. This caused the constant rate of decrease in SpO_2 in our model, whereas in the data we observed that the decrease depends on the apnea length. This constant slope in the model is explained by the choice that metabolic consumption was assumed independent of apnea length. This result was found since the threshold $CO_{2,th}$ was not reached during our simulations. In case this threshold was exceeded oxygen consumption would have dropped, directly affecting the slope of decrease as shown for apneas extracted from our patients.

The influence of the setting of other parameters, namely RR, Q_m and C_{AI}, on the simulated apneas was investigated. RR determined the steady state level of arterial SpO_2 and affected the rate at which this level dropped during early apnea. Q_m directly determined the rate of decrease in SpO_2 during late apnea. C_{AI}, an important parameter since low values are associated with a higher risk of respiratory distress syndrome, showed very limited contributions on the simulated apneas. Since the estimation of Q_m (both the

value and its independence on apnea length) largely influences the extent of the decrease in SpO_2, improving the estimation based on clinical data could improve apnea simulation.

Following a simulated apnea, the model showed a slight raise in HR, caused by the response to the baroreflex regulation. Our model was not able to simulate the observed drop in HR following CASE-Ds in our data. The modeled increase in HR was due to the increase of the intrathoracic pressure during apnea, resulting in a decrease in cardiac filling. Following the Starling effect [38] this led to a decrease in cardiac output and arterial pressure. However, this rise in HR was very small (< 0.8 bpm), and thus physiologically irrelevant. To explain the large decrease in HR found in our experimental data, our model should be extended with additional regulatory mechanisms, responding to inputs from the chemoreceptors and the lung stretch receptors. Notably, some studies claimed that the fall in HR is a chemoresponse to the fall in SpO_2 resulting from apnea, because of the significant correlation between the decrease in SpO_2 and HR [39]. However, our data suggests that the fall in HR precedes the fall in SpO_2, making this mechanism less plausible.

Results provided in this work allowed to compare CASE-Ds extracted from physiological data of premature infants with simulations of apnea from a cardiorespiratory model. These results shared different similarities including a far deeper decrease in SpO_2, where the minimum was reached later in time, in case of longer apneas. Furthermore, in both CASE-Ds and apnea simulations the initial SpO_2 levels are reached earlier after the minimum has been reached in case of shorter events. Some differences between the results could be explained either looking at how measures are performed in clinical practice in our neonatal intensive care unit (e.g., delayed decline in SpO_2 after apnea onset) and considering the limited number of recorded CASE-Ds within a certain group (e.g., CASE-Ds longer than 80 s) and the presence of short CASEs which often preceded CASE-Ds extracted from our dataset of premature infants. In addition to these, the slope of decrease in SpO_2 during simulated apneas was found to be the same for each apnea length since the simulated conditions of the infant before each apnea were the same for every situation, a condition that is not found in real patients.

Finally, an important limitation is the missing contribution of chemoreceptors and lung stretch receptors, that might affect the metabolic rate Q_m, which was assumed to remain constant during apnea during our simulations. Future studies should also try to model the fall in HR and possibly get us a better understanding of the underlying mechanisms that cause this fall, taking into account the contribution from the chemoreceptor, similarly to what was described by Albanese et al. [17]. Another limitation of our model is that it only allows to simulate central apneas while at the moment other types of apneas

cannot be simulated. Nevertheless, our cardiorespiratory model was successfully able to replicate the steady state of premature infants and describe the decay in SpO_2 during apnea, showing results in line with different literature studies.

6.5. Conclusion

The aim of this study was to combine the use of modeling data and real infant data to enhance understanding of AOP and as such, confirm dependencies between variables known to influence AOP. By analyzing CASE-Ds in real infant data, we showed that longer cessations of breathing were followed by a slower decrease in HR and SpO_2. The decrease in HR was also found to occur before the decrease in SpO_2. By extending an existing cardiovascular model to include a lung mechanics and a gas exchange compartment and by fitting parameters for a preterm infant, we were able to simulate different characteristics found in CASE-Ds (i.e., the drop in SpO_2 during apneas as well as a deeper decline in SpO_2 for longer apneas). Our model did not allow yet to simulate the decrease in HR, an observation which might be reproduced by introducing the chemoreceptors and lung stretch receptors in the model.

References

[1] N. N. Finer, R. Higgins, J. Kattwinkel, and R. J. Martin, "Summary proceedings from the apnea-of-prematurity group," *Pediatrics*, vol. 117 (Suppl, no. 3, pp. S47–S51, 2006.

[2] J. M. Abu-Shaweesh and R. J. Martin, "Neonatal apnea: What's new?," *Pediatr. Pulmonol.*, vol. 43, no. 10, pp. 937–944, 2008.

[3] C. F. Poets *et al.*, "Association between intermittent hypoxemia or bradycardia and late death or disability in extremely preterm infants," *JAMA - J. Am. Med. Assoc.*, vol. 314, no. 6, pp. 595–603, 2015.

[4] T. Carbone, L. C. Marrero, J. Weiss, M. Hiatt, and T. Hegyi, "Heart Rate and Oxygen Saturation Correlates of Infant Apnea," *J. Perinatol.*, vol. 19 (1), pp. 44–47, 1999.

[5] M. A. Mohr *et al.*, "Very long apnea events in preterm infants," *J. Appl. Physiol.*, vol. 118, no. 5, pp. 558–568, 2015.

[6] S. A. Sands, B. A. Edwards, V. J. Kelly, M. R. Davidson, M. H. Wilkinson, and P. J. Berger, "A model investigation of the impact of ventilation-perfusion mismatch on oxygenation during apnea in preterm infants," *J. Theor. Biol.*, vol. 264, no. 3, pp. 657–662, 2010.

[7] O. P. Mathew, "Apnea of prematurity: Pathogenesis and management strategies," *J. Perinatol.*, vol. 31, no. 5, pp. 302–310, 2011.

[8] J. M. Di Fiore, R. J. Martin, and E. B. Gauda, "Apnea of prematurity - Perfect storm," *Respir. Physiol. Neurobiol.*, vol. 189, no. 2, pp. 213–222, 2013.

[9] J. L. Feldman, C. A. Del Negro, and P. A. Gray, "Understanding the rhythm of breathing: So near, yet so far," *Annu. Rev. Physiol.*, vol. 75, pp. 423–452, 2013.

[10] H. Lee *et al.*, "A new algorithm for detecting central apnea in neonates," *Physiol. Meas.*, vol. 33, no. 1, pp. 1–17, 2012.

[11] W. Jennekens, M. Dat, P. H. M. Bovendeerd, P. F. F. Wijn, and P. Andriessen, "Validation of a preterm infant cardiovascular system model under baroreflex control with heart rate and blood pressure data," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 896–899, 2011.

[12] G. Varisco *et al.*, "Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU," *Acta Paediatr. Int. J. Paediatr.*, vol. 110, no. 4, pp. 1141–1150, 2021.

[13] C. D. Sá Couto, W. L. Van Meurs, J. A. Goodwin, and P. Andriessen, "A Model for Educational Simulation of Neonatal Cardiovascular Pathophysiology," *Simul. Healthc.*, no. 1, pp. 4–12, 2006.

[14] J. Wesseling and K. Settels, *Circulatory model of baro-and cardio-pulmonary reflexes*. 1993.

[15] M. Ursino, "Interaction between carotid baroregulation and the pulsating heart: A mathematical model," *Am. J. Physiol. - Hear. Circ. Physiol.*, vol. 275, no. 5 44-5, pp. 1733–1747, 1998.

[16] M. Ursino and E. Magosso, "Role of short-term cardiovascular regulation in heart period variability: A modeling study," *Am. J. Physiol. - Hear. Circ. Physiol.*, vol. 284, no. 4 53-4, pp. 1479–1493, 2003.

[17] A. Albanese, L. Cheng, M. Ursino, and N. W. Chbat, "An integrated mathematical model of the human cardiopulmonary system: Model development," *Am. J. Physiol. - Hear. Circ. Physiol.*, vol. 310, no. 7, pp. H899–H921, 2016.

[18] W. E. L. Brown and A. V. Hill, "The Oxygen-Dissociation Curve of Blood.," *R. Soc.*, vol. 94 (661), pp. 297–334, 1922.

[19] M. Delivoria-Papadopoulos, N. P. Roncevic, and F. A. Oski, "Postnatal changes in oxygen transport of term, premature, and sick infants: The role of red cell 2,3-diphosphoglycerate and adult hemoglobin," *Pediatr. Res.*, vol. 5, no. 6, pp. 235–245, 1971.

[20] K. Bauer, C. Uhrig, P. Sperling, K. Pasel, C. Wieland, and H. T. Versmold, "Body temperatures and oxygen consumption during skin-to-skin (kangaroo) care in stable preterm infants weighing less than 1500 grams," *J. Pediatr.*, vol. 130, no. 2, pp. 240–244, 1997.

[21] R. M. Kliegman, B. Stanton, J. S. Geme, N. Schor, and R. E. Behrman, *Nelson Textbook of Pediatrics*. 2011.

[22] K. Charles and S. Istvan, *Hemodynamics and Cardiology: Neonatology Questions and Controversies*. 2012.

[23] A. Castillo *et al.*, "Pulse oxygen saturation levels and arterial oxygen tension values in newborns receiving oxygen therapy in the neonatal intensive care unit: Is 85% to 93% an acceptable range?," *Pediatrics*, vol. 121, no. 5, pp. 882–889, 2008.

[24] S. P. Wardle and A. M. Weindling, "Peripheral oxygenation and anemia in preterm babies," *Pediatr. Res.*, no. 44 (1), pp. 125–131, 1998.

[25] P. Latzin *et al.*, "Lung volume, breathing pattern and ventilation inhomogeneity in preterm and term infants," *PLoS One*, vol. 4, no. 2, 2009.

[26] H. Suga, "Total mechanical energy of a ventricle model and cardiac oxygen consumption.," *Am. J. Physiol.*, vol. 236, no. 3, 1979.

[27] H. Suga, R. Hisano, and S. Hirata, "Heart rate-independent energetics and systolic pressure-volume area in dog heart," *Am. J. Physiol. - Hear. Circ. Physiol.*, vol. 13, no. 2, 1983.

[28] R. Joshi, C. Van Pul, L. Atallah, L. Feijs, S. Van Huffel, and P. Andriessen, "Pattern discovery in critical alarms originating from neonates under intensive care," *Physiol. Meas.*, vol. 37, no. 4, pp. 564–579, 2016.

[29] K. Bauer, O. Linderkamp, and H. T. Versmold, "Systolic blood pressure and blood volume in preterm infants," *Arch. Dis. Child.*, vol. 69, pp. 521–522, 1993.

[30] N. Aladangady, S. McHugh, T. C. Aitchison, C. A. J. Wardrop, and B. M. Holland, "Infants' blood volume in a controlled trial of placental transfusion at preterm delivery," *Pediatrics*, vol. 117, no. 1, pp. 93–98, 2006.

[31] K. H. Hsu, T. W. Wu, Y. C. Wang, W. H. Lim, C. C. Lee, and R. Lien, "Hemodynamic reference for neonates of different age and weight: A pilot study with electrical cardiometry," *J. Perinatol.*, vol. 36, no. 6, pp. 481–485, 2016.

[32] A. L. Kent, S. Meskell, M. C. Falk, and B. Shadbolt, "Normative blood pressure data in non-ventilated premature neonates from 28-36 weeks gestation," *Pediatr. Nephrol.*, vol. 24, no. 1, pp. 141–146, 2009.

[33] V. Soloveychik, A. Bin-Nun, A. Ionchev, S. Sriram, and W. Meadow, "Acute hemodynamic effects of caffeine administration in premature infants," *J. Perinatol.*, vol. 29, no. 3, pp. 205–208, 2009.

[34] N. N. Finer, K. J. Barrington, B. J. Hayes, and A. Hugh, "Obstructive, mixed, and central apnea in the neonate: Physiologic correlates," *J. Pediatr.*, vol. 121, no. 6, pp. 943–950, 1992.

[35] E. C. Fletcher, R. Kass, J. I. Thornby, J. Rosborough, and T. Miller, "Central venous O2 saturation and rate of arterial desaturation during obstructive apnea," *J. Appl. Physiol.*, vol. 66, no. 3, pp. 1477–1485, 1989.

[36] N. Iwase *et al.*, "Effects of Repetitive Airway Obstruction on O2 Saturation and Systemic and Pulmonary Arterial Pressure in Anesthetized Dogs," *Am. Rev. Respir. Dis.*, 1992.

[37] M. H. Wilkinson, P. J. Berger, N. Blanch, and V. Brodecky, "Effect of venous oxygenation on arterial desaturation rate during repetitive apneas in lambs," *Respir. Physiol.*, vol. 101, no. 3, pp. 321–331, 1995.

[38] A. M. Katz, "Ernest Henry Starling, his predecessors, and the 'Law of the Heart'.," *Circulation*, vol. 106, no. 23, pp. 2986–2992, 2002.

[39] D. J. Henderson-Smart, M. C. Butcher-Puech, and D. A. Edwards, "Incidence and mechanism of bradycardia during apnoea in preterm infants," *Arch. Dis. Child.*, vol. 61, no. 3, pp. 227–232, 1986.

[40] A. C. Guyton and J. E. Hall, *Textbook of medical physiology*, 11th ed. 2006.

Appendix 6.1: Lung mechanics model

The lung mechanics model used in this work is based on the work from Albanese et al. [17] and is outlined in **Figure 6.1B**. This model, which is used to describe pressures, flows and volumes by representing resistances and compliances by using segments and nodes respectively, is based on the equivalence with a resistor–capacitor circuit. To model the respiration, a pressure change in $p_{mus}(t)$ caused by the respiratory muscles is used. This pressure is described by taking into account the minimum end-inspiratory pressure $p_{mus,min}$, inspiration and expiration times T_I and T_E as well as the complete respiratory period T and by setting the time constant of the exponential expiratory profile $\tau = 0.2 \cdot T_E$:

$$p_{mus}(t) = \begin{cases} \frac{-p_{mus,min}}{T_I \cdot T_E} \cdot t^2 + \frac{p_{mus,min} \cdot T}{T_I \cdot T_E} \cdot t & t \in [0, T_I] \\ \frac{p_{mus,min} \cdot T}{1 - e^{T_E/\tau}} \cdot \left(e^{t - T_I/\tau} - e^{-T_E/\tau}\right) & t \in [T_I, T] \end{cases}$$

Changes in the pressures cause an air flow at each step, leading to changes in volume for each compartment. Alveolar pressure p_{Al} is the result of the following equation:

$$p_{Al} = p_{pl} + \frac{V_{Al} + V_{0,Al}}{C_{Al}}$$

Where p_{pl} indicates the pleural pressure and $V_{0,Al}$ the unstressed alveolar volume. p_{pl} as well as dead space pressure p_{ds} can be deducted in a similar way. The pressures induce air flows, depending on the airway resistances. Flow from bronchi to alveoli Q_{bAl}, which is equal to change in alveolar volume dV_{Al}, can be described by the following equation:

$$q_{bAl} = dV_{Al} = \frac{p_{ds} - p_{Al}}{R_{bAl}}$$

where R_{bAl} indicates the resistance in the segment connecting bronchi and alveoli. Following this equation q_{ds} as well dV_{ds} and dV_{pl} can also be extracted.

Appendix 6.2: Oxygen model

The amount of oxygen that is stored in a node depends on the blood volume V and the oxygen concentration cO_2:

$$VO_2 = cO_2 \cdot V$$

cO_2 only depends on the amount of oxygen bound to hemoglobin Hb, since the contribution of dissolved oxygen is very small and is therefore neglected:

$$cO_2 = \alpha \cdot Hb \cdot SpO_2$$

where Hb indicates the hemoglobin concentration and α the maximum binding capacity of Hb. The oxygen saturation SpO_2 is dependent on the partial oxygen pressure pO_2 according to the Hill equation [18]. This equation relates to the oxy-hemoglobin dissociation curve of a premature infant since pO_2^{50} is the oxygen pressure at a saturation of 50%:

$$SpO_2 = \frac{(pO_2/pO_2^{50})^n}{1 + (pO_2/pO_2^{50})^n}$$

The amount of oxygen VO_2 in a node changes due to convective transport Q_C, diffusive transport Q_D, and metabolic consumption Q_m in the systemic arterial micro-circulation, according to:

$$\frac{dVO_2}{dt} = Q_{c,in} - Q_{c,out} + Q_D - Q_m$$

Q_C depends on the flow q through a node and the oxygen concentration cO_2:

$$Q_c = q \cdot cO_2$$

Q_D depends on the diffusion capacity D_{lung}, as shown in the following equation:

$$Q_D = D_{lung} \cdot (pO_{2,air} - pO_{2,blood})$$

To estimate D_{lung} steady state conditions are assumed and the oxygen balance for the blood volume as a whole is considered, implying that the total diffusive uptake is equal to the total metabolic consumption. Q_C takes place in the dead space and alveoli and Q_D only occurs from the alveoli, giving the following relations for oxygen volumes VO_2 in the dead space and alveoli:

$$\frac{dVO_{2,ds}}{dt} = Q_{cds,in} - Q_{cds,out}$$

$$\frac{dVO_{2,Al}}{dt} = Q_{cAl,in} - Q_{cAl,out} - Q_D$$

Oxygen pressures in the compartments can be determined with oxygen concentrations and atmospheric pressures, while correcting for the pressure of the water vapour p_{vap} as the air is humidified as it enters the lungs [40]:

$$pO_{2,Al} = cO_{2,Al} \cdot (p_{atm} - p_{vap}) \qquad with\ cO_{2,Al} = \frac{VO_{2,Al}}{V_{Al}}$$

$$pO_{2,ds} = cO_{2,ds} \cdot (p_{atm} - p_{vap}) \qquad with\ cO_{2,ds} = \frac{VO_{2,ds}}{V_{ds}}$$

Oxygen concentrations are derived from the ratio of oxygen and air volume in a compartment.

As described by Sa Couto et al. [13], Q_m is set to a constant value unless it is below a certain threshold level, where Q_m depends on the concentration level in the blood:

$$Q_m = \begin{cases} Q_{m,0} & if\ cO_2 \geq cO_{2,th} \\ Q_{m,0} + K_f(cO_2 - cO_{2,th}) & if\ cO_2 < cO_{2,th} \end{cases}$$

with $cO_{2,th}$ being the threshold concentration, K_f a constant and $Q_{m,0}$ the normal metabolic rate.

Appendix 6.3: Parameters for the lung mechanics and oxygen model

A

Parameter	Description	Value
RR	Respiration rate	49 bpm
$p_{mus,min}$	Minimal pressure of resp. muscles	-3 mmHg
$C_{Al,tot}$	Alveolar compliance	0.93 ml/cmH_2O
C_{ds}	Dead space compliance	0.07 ml/cmH_2O
C_{cw}	Chest wall compliance	5.006 ml/cmH_2O
R_{ds}	Dead space resistance	78.9 cmH_2O • ms/ml
$R_{bAl,tot}$	Bronchi to alveoli resistance	3.68 cmH_2O • ms/ml
FRC	Functional residual capacity	23.24 ml
$V_{0,ds}$	Unstressed dead space volume	2.03 ml
$V_{0,Al,tot}$	Unstressed alveoli volume	5.0 ml

B

Parameter	Description	Value
$cO_{2,ao}$	Atmospheric oxygen concentration	0.21 ml O_2/ml air
SpO_2^{50}	Oxygen pressure at 50% saturation	19 mmHg
α	Hemoglobin affinity	1.40 g Hb/ml blood

(A) Parameters for the lung mechanics and (B) for the oxygen model. In addition to the values presented here the parameters respiration time T_{breath}, time constant τ, inspiration/expiration ratio IE and pressure at airway opening p_{ao} were defined in agreement with Albanese et al. following the definition of the respiration rate RR [17]. For the oxygen model, atmospheric pressure p_{atm}, hemoglobin concentration Hb were defined as in the work from Albanese et al. [17], constant n as in the work from Hill et al. [18], the metabolic rate $Q_{m,0}$ as in the work from Bauer et al. [20] and the metabolic constant K_f and threshold concentration for metabolism $cO_{2,th}$ as in the work from Sa Couto et al. [13].

CHAPTER 7

Two-step data-driven central apnea detection using machine learning

Reproduced after: Central apnea detection in premature infants using machine learning, Computer Methods and Programs in Biomedicine, 2022, 226, pp. 107155.

Abstract

Apnea of prematurity is one of the most common diagnoses in neonatal intensive care units. Apneas can be classified as central, obstructive or mixed. According to the current international standards, minimal fluctuations or absence of fluctuations in the chest impedance (CI) suggest a central apnea (CA). However, automatic detection of reduced CI fluctuations leads to a high number of central apnea- suspected events (CASEs), the majority being false alarms. We aim to improve automatic detection of CAs by using machine learning to optimize detection of CAs among CASEs.

Using an optimized algorithm for automated detection, all CASEs were detected in a population of 10 premature infants developing late-onset sepsis and 10 age-matched control patients. CASEs were inspected by two clinical experts and annotated as CAs or rejections in two rounds of annotations. A total of 47 features were extracted from the ECG, CI and oxygen saturation signals considering four 30 s-long moving windows, from 30 s before to 15 s after the onset of each CASE, using a moving step size of 5 s. Consecutively, new CA detection models were developed based on logistic regression with elastic net penalty, random forest and support vector machines. Performance was evaluated using both leave-one-patient-out and 10-fold cross-validation considering the mean area under the receiver-operating-characteristic curve (AUROC).

The CA detection model based on logistic regression with elastic net penalty returned the highest mean AUROC when features extracted from all four time windows were included, both using leave-one- patient-out and 10-fold cross-validation (mean AUROC of 0.88 and 0.90, respectively). Feature relevance was found to be the highest for features derived from the CI. A threshold for the false positive rate in the mean receiver-operating-characteristic curve equal to 0.3 led to a high percentage of correct detections for all CAs (78.2%) and even higher for CAs followed by a bradycardia (93.4%) and CAs followed by both a bradycardia and a desaturation (95.2%), which are more critical for the well-being of premature infants.

Models based on machine learning can lead to improved CA detection with fewer false alarms.

7.1. Introduction

Apnea of prematurity is a condition that is commonly found in premature infants admitted to neonatal intensive care units (NICUs) and is related to the immaturity of their respiratory control system [1]. Patients who develop severe illnesses such as sepsis or necrotizing enterocolitis are particularly prone to recurrent apneas [2]–[4]. Furthermore, apnea of prematurity is associated with later neurodevelopmental impairment and brain damage [5], [6]. Accurate and timely apnea detection is therefore clinically important to improve outcomes in premature infants.

Apnea of prematurity is most often defined as a cessation of breathing lasting more than 20 s or more than 10 s when followed by a bradycardia, desaturation, or both [7], [8]. However, some clinical experts disagree on the second part of the definition as they consider it necessary to have at least 15 s of cessation of breathing to define an apnea, even when followed by bradycardia or desaturation [9]–[11].

Apneas can be categorized as central apnea (CA), characterized by a cessation of respiratory drive with a lack of diaphragmatic activity, obstructive apnea, characterized by an airway obstruction which causes an absence of airflow, and mixed apnea, which includes characteristics of both central and obstructive apnea [1], [12]–[14]. The heart rate (HR) and oxygen saturation (SpO_2) responses vary depending both on the different apnea types as well as their duration [11]. For instance, it was noticed that a fall in HR to less than 100 beats/min is typically found in the case of CAs whereas an initial acceleration in the HR is more often present during obstructive apneas [11]. At the same time, further decreases in SpO_2 were noticed in longer apneas, irrespectively of their type [11], [15].

Premature infants are constantly monitored, and an alarm is generated by the patient monitor in the event of an apnea. In current clinical practice, patient monitors generate an apnea alarm that is based on temporary reduction in amplitude of the chest impedance (CI), since this reflects a reduction in chest motion and thus indirectly in breathing motion [1], [13]. This approach leads to the identification of the so-called central apnea-suspected events (CASEs), which may reflect apneic events of central origin. In fact, while among CASEs there are true CAs, several CASEs are also known to be false alarms. Studies have shown the presence of a high number of false apnea associated alarms in the daily NICU practice [9], [16]–[19], indicating that the current clinical apnea detection is deficient.

In the past years, different strategies have been developed with the aim to improve apnea detection, encompassing a variety of signal processing techniques as well as

machine learning and deep learning techniques [17], [20]–[25]. Among the different proposed solutions, the algorithm proposed by Lee et al. focuses specifically on CA detection [17]–[19]. In a previous study, we optimized this algorithm considering annotations from two clinical experts [26]. We observed a very high recall (sensitivity or true positive rate (TPR)) for CAs, but a persistently low precision (or positive predictive value (PPV)). This indicated that many false CAs were still detected. In order to obtain both a high precision and a high recall, the aim of the current study was to develop an optimized detection model based on machine learning to automatically detect true CAs from all CASEs extracted using the optimized algorithm for CA detection [26].

7.2. Methods

7.2.1. Dataset

We used routinely obtained data from premature infants that participated in a previous matched case-control study on sepsis [27]. These were all patients admitted to the NICU of Máxima Medical Center in Veldhoven, the Netherlands from July 2016 to December 2018. From that existing database we selected premature infants most likely to experience apneas (i.e. all very preterm infants with a gestational age ≤ 30 weeks) that developed late-onset sepsis (LOS), and their case-matched controls.

LOS was diagnosed when patients presented the clinical signs of a generalized infection according to the Vermont Oxford Criteria [28], required the use of antibiotics after at least 72 hours of life, and pathogens were isolated from their blood culture. For these patients we identified the CRASH-moment (blood Culture collection, Resuscitation, and Antibiotics Started Here), following the definition originally provided by Griffin and Moorman [29]. We then collected all the data for all 48 hours prior to the onset of sepsis, in which apneas are very likely to occur [3], [26], [30]. For the case-matched controls we calculated an equivalent-in-time CRASH-moment to make sure that the postmenstrual age was matched and we extracted the data from the 48 hours prior to the equivalent CRASH-moment. A total of 960 hours of data from 20 premature infants (10 LOS patients and 10 matched controls) were included in this study. All patients' characteristics, specified for the LOS and control groups, are presented in **Table 7.1**.

Each patient was monitored using Philips IntelliVue MX800 patients' monitors (Philips Medical Systems, Böblingen, Germany) according to the present clinical standard and all data was stored in a data warehouse (PIIC-iX, Data Warehouse Connect; Philips Medical Systems, Andover, MA). The ECG and CI were measured using three ECG leads with a sampling frequency of 250 Hz for the ECG and 62.5 Hz for the CI. The SpO_2 was measured

with a photoplethysmogram with a sampling frequency of 1 Hz. All these three signals were then extracted from the data warehouse for further processing.

As this study had a retrospective and non-invasive nature, a waiver was provided by the medical ethical committee in accordance with the Dutch law on medical research with humans (WMO).

7.2.2. Annotation process

As a first step in this study, we applied the optimized algorithm for CA detection described in our previous chapter [26] on our dataset of 20 premature infants. This algorithm was derived from an algorithm originally described by Lee et al. [17]–[19], which allows to filter out the cardiac artifacts found in the ECG from the CI, to compute a CA probability function from the resulting filtered respiration signal and finally to detect CASEs, which include both CAs and false CAs. We previously optimized this algorithm to improve its precision without affecting its high recall [26].

In the current study, the optimized algorithm for CA detection was applied to the patients' signals to detect CASEs longer than 10 s, in a similar way to the previous use of this algorithm [26]. This first step is indicated in **Figure 7.1**, which also includes an overview of all subsequent steps performed in this study. Matlab (The MathWorks, Natick, United States) was used to detect CASEs from our dataset, as well as to perform the annotation process and the signal processing and feature extraction techniques.

Table 7.1: Characteristics of the patients included in the current study, separated considering the Late-Onset Sepsis (LOS) and control groups and all patients together. Count of CASEs and CAs, in particular, refer to those used in the machine learning experiments.

Dataset	LOS	Control	All
Count of patients	10	10	20
Gestational age (weeks) (mean ± std)	27.9 ± 0.9	27.9 ± 1.0	27.9 ± 0.9
Postnatal days (mean ± std)	10 ± 8	10 ± 9	10 ± 8
Postmenstrual age (weeks) (mean ± std)	29.3 ± 0.9	29.3 ± 0.9	29.3 ± 0.9
Birth weight (g) (mean ± std)	998 ± 216	1091 ± 156	1044 ± 190
Gender (female)	50%	50%	50%
Hours per patient	48	48	48
Count of CASEs (median [IQR])	351.5 [284.3 – 390.5]	231.5 [156.5 – 300.5]	284.5 [187.5 – 367]
Count of CAs (median [IQR])	47.5 [18.7 – 76.5]	17 [4.2 – 44]	26.5 [9.5 – 62.2]

A total of 7818 CASEs were extracted from the 20 premature infants. Two clinical experts were asked to provide, independently from each other, annotations for all the extracted CASEs using a user interface developed for this task. Three different options were given for the annotations: 'central apnea' (used for CASEs that clearly presented the characteristics of a CA as pre-defined by the clinical experts, i.e. a trace of flat CI following a previously regular fluctuating CI), 'rejection' (used when a CASE did not present the typical characteristics of a CA, i.e. in case of a flat CI trace mixed with large CI fluctuations), and 'error' (used in case of missing or corrupt signal).

A second (final) round of annotations was performed with the two clinical experts present at the same time to find a consensus for all CASEs that presented different annotations after the first round.

7.2.3. Time windows for CASE detection

We observed a high number of clusters of CASEs, in which CASEs were present at a brief distance from other CASEs. In particular, 2172 CASEs were found to be preceded by another CASE in the previous 30 s (27.8% of the total count of CASEs). We therefore performed a cluster analysis to investigate whether patterns in the transitions for the annotations from the clinical experts were present. We investigated the likelihood of having the same annotation in subsequent CASEs, information that could possibly be relevant in future detection or prediction studies.

Only CASEs which were not preceded by another CASE in the previous 30 s were included to perform machine learning experiments. This clinical choice was motivated by the fact that the decrease in HR or oxygen saturation due to a previous CASE might affect the vital signs in the patient monitor and the derived feature values computed for a subsequent CASE. This could have therefore possibly caused a wrong detection and influenced the learning process of the detection models based on machine learning. Previous studies indicated that the definition of 7.5-min pre-apnea (as well as post-apnea) windows might be an optimal solution to perform a prediction study for apnea of prematurity [20], [24], [31] but also suggested that reduction of this window length would better suit the needs of the clinical practice. Our choice yielded a total of 5646 CASES to be included in (72.2% of the initial CASEs).

After this process, all the remaining CASEs having an 'error' annotation were discarded from the study. This resulted in a total of 5460 CASEs to be considered for the upcoming steps (69.8% of the initial CASEs).

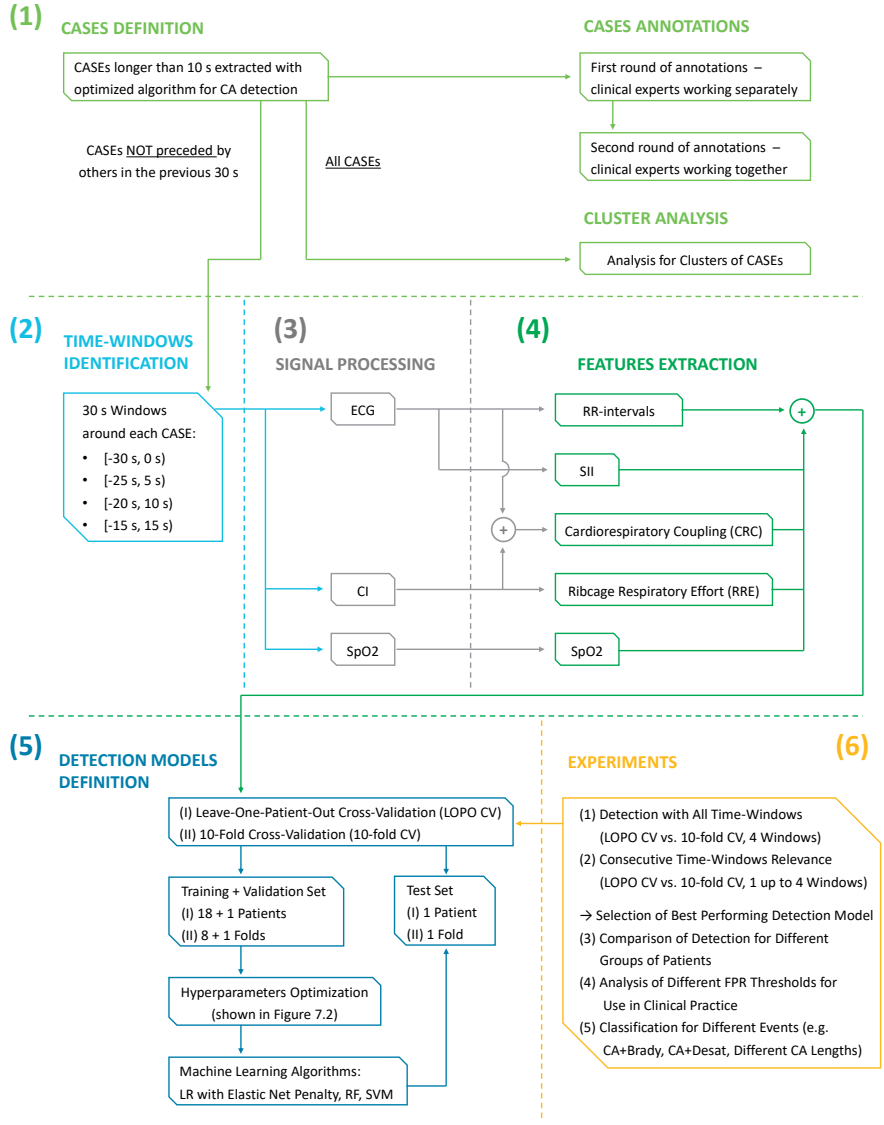


Figure 7.1: Overview of the steps performed in this study starting from the extraction of central apnea-suspected events (CASEs) and their annotation by clinical experts (1). Next, CASEs cluster analysis was performed and for each CASE not preceded by another CASE in the previous 30 s, four 30-s windows with 5 s overlap ([-30s, 0s), [-25s, 5s), [-20s, 10s) and [-15s, 15s)) were identified (2). Signals extracted from the patients were then processed to create five new signals (3). From all these new signals 47 features were extracted (4). Two different cross-validation methods, leave-one-patient-out cross-validation (LOPO CV (I)) and a 10-fold cross-validation (10-fold CV (II)) and three different machine learning algorithms, logistic regression with elastic net penalty (LR), random forest (RF) and support vector machines (SVM), were defined to create detection models at step (5). Finally, five different machine learning experiments were defined and performed (6).

For all the included CASEs, we included information starting from 30 s before the start of a CASE (-30 s) to 15 s after the start of the CASE and we used a moving step size of 5 s. This time window was chosen so that the included time was shorter than the minimum time requested by different apnea definitions to provide an annotation (15 s of cessation of breathing) [9]–[11]. This criterion led to feature extraction from four 30-s windows: [-30 s, 0 s), [-25 s, 5 s), [-20 s, 10 s) and [-15 s, 15 s), where 0 s was the moment of the start of a CASE.

7.2.4. Signal processing

RR-intervals were derived from the ECG by applying an algorithm for R-peaks detection and computing the time between consecutive R-peaks, using the same method described in our previous studies [27], [32]. From the ECG, the signal instability index (SII) was then calculated with the aim of quantifying patient motion [27], [32]. This measure was obtained by applying a band-pass filter (0.001–0.40 Hz) to 10 s-long ECG windows and subsequently by computing a kernel density estimate to return a measure of the patient motion for every second (1 Hz). Low values for this signal are associated with an absence of movement whereas higher values suggest that a movement is performed by the patient.

The CI was processed following the steps described by Redmond et al. to obtain the respiration Ribcage Respiratory Effort (RRE) [33]. This signal was computed by subtracting its mean and by removing the HF noise present above respiratory frequencies by applying a low-pass filter with a tenth-order Butterworth filter with a cutoff equal to 0.8 Hz. The resulting signal obtained for each patient received an amplitude normalization to remove the contribution of the tidal volume, a parameter that is different for each patient. In order to perform the normalization, turning points were first identified and the median peak-to-trough amplitude was found considering all differences between consecutive peaks and troughs. This value was then used to normalize the whole patient signal. The cardiorespiratory coupling signal (CRC) was finally computed by resampling the respiration RRE in correspondence with R-peaks in the ECG. No additional processing was performed on the SpO_2 measured with the photoplethysmogram.

7.2.5. Feature extraction

The RR-intervals, SII, respiration RRE, CRC and SpO_2 were all used to extract a total of 47 features for 5460 CASEs. Detailed information regarding all features can be found in **Appendix 7.1** and only features that are less commonly used are more extensively explained in this section, with our motivation to include them in our detection models. Each feature, unless stated differently, was extracted considering 30-s windows following the windows defined in **Paragraph 7.2.3**.

Features extracted from the RR-intervals using time-domain, frequency-domain and non-linear measures are well-established in previous studies. These include the Mean RR, the standard deviation of the normal RR intervals (SDNN) and the root mean square of successive differences between adjacent normal RR intervals (RMSSD), which can provide an estimate of overall heart-rate variability as well as an estimate of its short-term components [34], [35]. We also determined Sample Asymmetry of the RR-intervals, a measure of symmetry in the histogram of RR-intervals, which in previous studies showed clear differences before different illnesses (e.g., sepsis and systemic inflammatory response syndrome) compared to other instances [36]. Sample entropy, a measure able to quantify similarity for different sequences within a time-series [37], was also applied to the RR-intervals [38]. The Percentage of Decelerations and Standard Deviation of Decelerations were calculated since both proved to be effective in predicting the onset of sepsis [27], [39].

The phase-rectified signal averaging (PRSA) [40], [41] is capable of quantifying quasi-periodic oscillations in a physiological signal by filtering out artefacts and noise. We identified anchor points considering both increases in the signal (increase events) as well as decreases (decrease events). Windows of length equal to 40 time samples were set around the anchor points (i.e. $L = 20$), a choice motivated by the short time windows (30 s) used in this study. From these PRSA waveforms seven different features were extracted [42]. These included the Immediate Deceleration and Acceleration Response (IDR and IAR) as well as the Average Deceleration and Acceleration Responses (ADR and AAR), with adjusted parameters considering the length of the time windows used in this study. In addition to these features, the Slope of the Deceleration and Acceleration Response (SDR and SAR), which are able to quantify the rate of HR response, were also included. Finally, a new feature extracted from the PRSA was developed and included in this study. This was defined by dividing the difference between the count of anchor points from the PRSA with decrease events and with increase events by the count of RR-intervals and then multiplying this value by 100 (Anchor Points Decrease versus Increase).

The Fast Fourier Transform (FFT) was applied on the RR-intervals after resampling this signal at 250 Hz to extract the power spectral densities, allowing to measure the development of the cardio-respiratory coupling and the presence of respiratory sinus arrhythmia (RSA), as suggested in previous studies [43], [44]. To achieve this aim, we used the frequency ranges indicated by Indic et al. [43] and defined the Low Frequency (LF: 0.01-0.15 Hz), High Frequency (HF: 0.15-0.45 Hz), Super High Frequency (sHF: 0.45-0.7 Hz) and Ultra High Frequency (uHF: 0.7-1.5 Hz) ranges. The main reason for this choice compared to those provided in other studies [33], [44] was the possibility of using the same frequency ranges also for the FFT of the respiration RRE. In addition to these

frequency-domain features, we also extracted the LF/HF Ratio and the Mean Respiratory Frequency derived from the RR-intervals, defined as the frequency of maximum power in the frequency range 0.15–0.45 Hz [33].

Four features were extracted from the SII to quantify the movements of premature infants. These were the Mean SII, the Standard Deviation of the SII, Skewness of the SII and Interdecile Range of the SII. These were selected for their ability to identify motion patterns [27], [32], [45].

From the respiration RRE we extracted the Mean RRE, Standard Deviation of the RRE, Skewness of the RRE and Interdecile Range of the RRE. Sample Entropy of the RRE was computed with the same parameter choices described for the RR-intervals. Other time-domain features extracted from the respiration RRE include the Envelope Power, the Breath-by-Breath Correlation, the Breath Length Variation and the Time Domain Respiratory Frequency, features that proved to be particularly useful in classifying sleep stages. A detailed explanation for these features can be found in the study from Redmond et al. [33]. The Standardized Median of the Peaks and the Standardized Median of the Troughs were then computed as additional features to quantify the amplitude of the respiratory peaks and troughs [46]. With regards to features from the frequency domain, the power spectral densities were computed from the FFT of the respiration RRE by using the same LF, HF, sHF and uHF ranges previously mentioned for the case of the RR-intervals [43]. Similarly to the case of the RR-intervals, the frequency of maximum power in the frequency range 0.15–0.45 Hz was also computed considering the respiration RRE (Mean Respiratory Frequency derived from the respiration RRE) [33].

Cardiorespiratory features, which have proven to be particularly useful to identify sleep stages as well as cardiorespiratory phase-coupling during obstructive sleep apnea in adults, were introduced to verify whether they can be helpful in the context of apnea detection in premature infants [47], [48]. For their computation we applied the nonlinear visibility graph (VG), a method which allows for the description of a time-series based on geometric criteria [49], to the CRC, as suggested by Long et al. [47]. This method requires the calculation of the degree δ of all data points in the time-series, also known as nodes, based on their visibility on other data points. The Mean Degree of the nodes and the Degree Variation of the constructed VG network were then introduced as features to quantify the degree of complexity of the time-series. The Assortativity Coefficient was instead introduced to estimate whether nodes were more likely to connect with high-degree or low-degree nodes [47]. Similarly to the study by Long et al., we used 3.5 min windows for the computation of features, in order to capture enough changes in the RR-intervals. However, we set the upper time limit for these extended windows to match

exactly the end of each 30 s windows previously defined (i.e. the last 30 s of the 3.5 min windows corresponded to the 30 s windows used to extract other features).

Features extracted from the SpO_2 included its mean and standard deviation. In addition, we also estimated the slope of linear fitting of all data points within a time window. These features were selected to provide an indication about whether a desaturation was present in the current time window or expected in the upcoming ones.

Finally, Z score normalization was applied to the feature matrix which included features extracted for each CASE considering the different time windows. After the feature extraction process, CASEs presenting features with two or more consecutive 'NaN' values were discarded resulting in a total of 5333 CASEs to be considered for the detection models (68.2% of the initial CASEs). As it is reported in **Table 7.1**, for each patient we found a median and [IQR] value of 284.5 [187.5 – 367] CASEs and 26.5 [9.5 – 62.25] CAs annotated by clinical experts in the 48 hours of data. In 5 patients very few CAs were annotated (count of CAs ≤ 5) while a high number of CAs was found in 6 other patients (count of CAs ≥ 50).

7.2.6. Machine learning: detection models

Detection models based on three traditional machine learning algorithms were employed in this study to perform CA detection: logistic regression (LR), random forest (RF) and support vector machines (SVM). LR was selected for its simplicity, which guarantees an easy transfer of knowledge and interpretability of the results. RF and SVM were selected following previous studies which proved that they can be suitable for apnea prediction in premature infants [22], [23].

LR was implemented with the elastic net penalty, which linearly combines the L1 and L2 penalty functions of the lasso (least absolute shrinkage and selection operator) and ridge methods. Elastic net was adopted since different studies highlighted that it usually holds a better predictive power compared to the lasso regression while still performing feature selection, a characteristic that was considered relevant for the relatively high number of included features compared to the number of CASEs (records in machine learning terminology) [50]. We performed hyperparameter optimization by means of cross-validation, further discussed in the next paragraph, for the parameter C, which indicates how much LR is penalized and L1 ratio, which influences the algorithm penalty choosing among a mere L1 penalty, a mere L2 penalty or a combination of L1 and L2 penalties.

Three hyperparameters were optimized for the RF: the number of trees, the number of candidate features to consider for the best split at each node, and the maximum depth of the tree.

Table 7.2: List of all machine learning algorithms and parameters values investigated to perform hyperparameter optimization for the different detection models.

Machine learning algorithms	Parameters	Parameters values
Logistic regression with elastic net penalty (LR)	C	0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1
	L1 ratio	0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1
Random forest (RF)	Number of trees	1, 2, 5, 10, 20, 50, 100, 200
	Number of features for the best split	1, 2, 5, 10, 'log2' (i.e. $\log_2 num.\ features$), 'auto' (i.e. $\sqrt{num.\ features}$)
	Maximum depth of the tree	1, 2, 5, 10, 'None' (i.e. nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples)
Support vector machines (SVM)	Regularization parameter	0.1, 0.5, 1, 5, 10, 50, 100
	Type of kernel	'rbf' (i.e. radial basis function), 'poly', 'sigmoid'
	Kernel coefficient	'scale' (i.e. 1/(num. features · var(input matrix))), 1, 0.1, 0.01, 0.001, 0.0001

Finally, three hyperparameters were optimized for the case of the SVM. These included the regularization parameter, the type of kernel to be used and kernel coefficient, suitable for all the different types of kernels used in this study. A list of all the values for each hyperparameter that were investigated is presented in **Table 7.2**. All the detection models here described were implemented using the Scikit-learn in Python (Python Software Foundation, Fredericksburg, United States).

7.2.7. Machine learning: validation and evaluation

Two cross-validation methods were used for each detection model to optimize the hyperparameters, as shown in **Figure 7.1**. The first cross-validation method was a leave-one-patient-out cross-validation (LOPO CV), using for each iteration the CASEs from one patient as a test set. LOPO CV was selected as it relates well to clinical practice, where each new patient admitted in a NICU is treated as a test set. CASEs from the remaining 19 patients were also separated using a second LOPO CV, which allowed us to obtain for each iteration 18 patients in the training set and 1 patient in the validation set in order to perform hyperparameter optimization for each detection model. All steps regarding hyperparameter optimization are shown in **Figure 7.2**. A grid search approach was used to determine the training and validation scores for each iteration of the second LOPO CV, computed by using the area under the receiver operating characteristic curve (AUROC) as a scoring function. The mean training scores, mean validation scores and the difference between the mean training and mean validation scores were computed

for each iteration of the first LOPO CV. The optimal set of parameters for the detection model to be used on a test set was then selected based on two requirements: a set of parameters leading to (1) a mean training score greater than a predefined threshold (Th_{class}), solution adopted to prevent underfitting, and to (2) the minimum difference between the mean training and mean validation scores, solution implemented to avoid overfitting. Th_{class} was empirically set equal to 0.9. In case less than 30% among the mean training scores computed for all sets of parameters were found to be greater than 0.9, its value was decreased each time by 0.05 and the full experiment repeated until this condition was fully satisfied for all the iterations of the first LOPO CV.

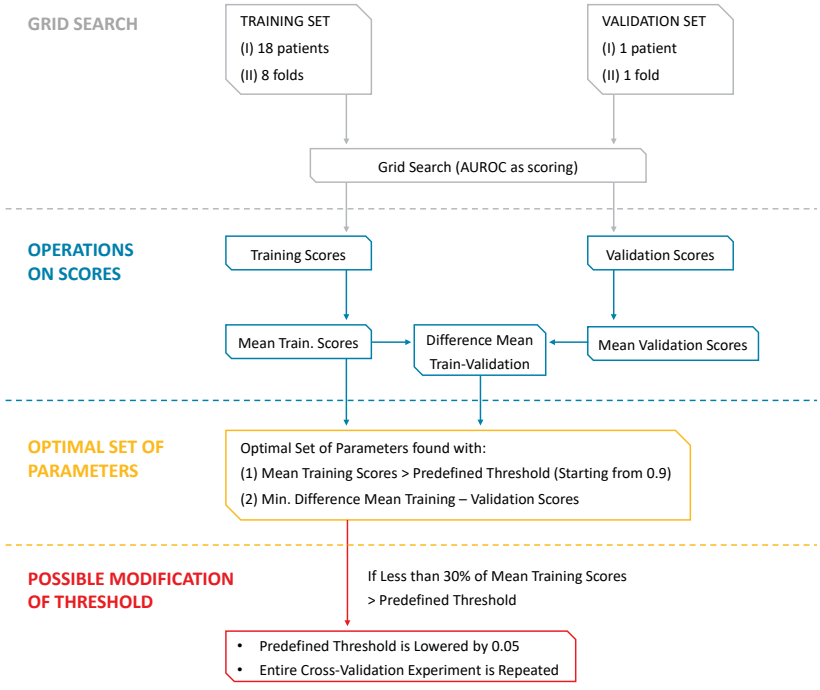


Figure 7.2: Steps used in this study to perform hyperparameter optimization for all detection models used in this study and for both cross-correlation methods, leave-one-patient-out cross-validation (LOPO CV (I)) and a 10-fold cross-validation (10-fold CV (II)). The optimal parameters found using these criteria are then applied to the patient/fold left as test set.

As the CAs were unevenly distributed over the patients, with 5 patients having only a few CAs, we also trained the detection models using a 10-fold cross-validation (10-fold CV), folded over CASEs rather than patients, as a second cross-validation method. This experiment allowed to distribute CAs more evenly across all the folds, resulting in each fold having around one CA per 6 CASEs. On the other hand, an important limitation of 10-fold CV is that the inclusion of CASEs from different patients in the training and test

sets can introduce overfitting, a problem that is not found with LOPO CV. Nonetheless, this cross-validation method was included in our study also to quantify the extent of the overfitting compared to LOPO CV. Furthermore, as shown in **Figure 7.2**, a second cross-validation consisting of 9 folds, a number selected to keep consistency with the choice applied for the second LOPO CV (where the number of patients was reduced by 1 compared to the first LOPO CV), was used to create the training and validation sets and to perform hyperparameter optimization, following the same steps as described above for LOPO CV.

7.2.8. Machine learning: experiments

Five different machine learning experiments were performed. These are also all indicated in **Figure 7.1**. All the experiments allowed to evaluate the detection of CAs among CASEs in the test set by using the mean AUROC as metric, unless stated otherwise.

As a first experiment, we evaluated the performance of the different detection models based on machine learning and two cross-validation methods on the detection of CAs using information from all time windows. Features extracted from all time windows [-30 s, 0 s), [-25 s, 5 s), [-20 s, 10 s) and [-15 s, 15 s) were fed to all detection models, for both cross-correlation methods. This allowed us to verify which detection model worked better for CA detection and whether AUROC values were consistent for both cross-correlation methods.

As a second experiment, we investigated the relevance of including features from an increasing number of time windows, starting from window [-30 s, 0 s) only and up to windows [-30s, 0s), [-25 s, 5 s), [-20 s, 10 s) and [-15 s, 15 s), on the detection. This experiment was also performed for all three detection models and both cross-validation methods.

Before proceeding with the following experiments, we identified the best performing detection model and investigated the feature relevance and the hyperparameters selected per each fold. However, only detection models created by means of LOPO CV were considered since they prevent overfitting (i.e. LOPO CV doesn't allow to use CASEs extracted from one patient both in the training and test set). In addition to this, the use of LOPO CV might be more reasonable in a clinical context, where an algorithm should work proficiently immediately after the admission of a new patient, where no training can be done with previous events extracted from this same patient. This detection model was then used to perform the remaining experiments.

The third experiment consisted of verifying whether the detection was performed differently for different groups of patients. At first, we compared AUROC values for

control and LOS patients and applied the Wilcoxon signed-rank test to compare AUROC values from paired patients belonging to the two groups. We then compared AUROC values for each patient experiencing a low number of CAs (count of annotated CAs ≤ 5) and high number of CAs (count of CAs ≥ 50). Due to the low number of patients in these two groups, for this comparison we only evaluated the mean, minimum and maximum AUROC for patients in the two groups.

As a fourth experiment we extracted the confusion matrices found after the definition of different thresholds for the false positive rate (FPR) in the mean ROC curve. This allowed to evaluate how different metrics, recall (TPR), precision (PPV), miss rate (or false negative rate (FNR)) and the specificity (or true negative rate (TNR)), varied per each selected FPR. This experiment allowed us to understand the feasibility of using the selected detection model and cross-validation method for performing CA detection in the clinical practice.

As a fifth and final experiment, after the identification of an optimal FPR, we investigated how the selected detection model performed in detecting different events (recall). At first, we investigated the percentages of correct detections (recall) for all CAs, CAs followed by a bradycardia (HR ≤ 80 bpm), CAs followed by a decrease in HR (80 bpm < HR ≤ 100 bpm), CAs followed by a desaturation (SpO_2 ≤ 80%), CAs followed by a decrease in SpO_2 (80% < SpO_2 ≤ 88%), and CAs followed by both a bradycardia and a desaturation (also often referred as apnea bradycardia desaturation events in literature [17], [19]). The definitions for bradycardia and desaturation were selected following the thresholds for alarms used in our NICU [51]. We then evaluated the percentages of correct detection of CAs with different length. For this purpose we separated CAs considering the following lengths L (i.e. 10 s ≤ L < 20 s, 20 s ≤ L < 30 s, 30 s ≤ L < 40 s, 40 s ≤ L < 50 s, 50 s ≤ L < 60 s, L ≥ 60 s) in a similar way to what was proposed by Mohr et al. for an analysis of different apnea lengths [19]. Finally, we investigated the percentages of correct detection for CASEs agreed and disagreed upon by clinical experts during the first round of annotations.

7.3. Results

Figure 7.3 shows results from the cluster analysis of CASEs, in which transitions for the annotations from the clinical experts were investigated. This analysis showed consistency for annotations from clinical experts within clusters, irrespectively of the transition that was considered. Considering the first transition as an example, clusters starting from a CA had a far higher likelihood of being followed by a CA rather than a rejection for the following CASEs (16.7% vs. 3.8%). The same consideration was found

for clusters starting from rejection, which were far more often followed by a rejection (24.4%) rather than a CA (1.1%). All the other percentages of transitions indicated instead the end of a cluster.

Our first machine learning experiment consisted of investigating the performance of the different detection models considering features from all time windows. **Figure 7.4** shows the mean ROC curves for the different detection models. **Figure 7.4A**, computed by using LOPO CV, shows that the highest AUROC was found when using the LR with elastic net penalty (0.88 ± 0.08). This result was followed closely by the AUROC found using SVM (0.87 ± 0.09) and RF (0.85 ± 0.09). Slightly higher results compared to those found with LOPO CV were found by means of 10-fold CV, as shown in **Figure 7.4B**, where the highest AUROC was found with LR with elastic net penalty (0.90 ± 0.02), result followed closely by the AUROC found with SVM (0.89 ± 0.02) and RF (0.87 ± 0.03).

Our second experiment investigated the relevance of including features from an increasing number of time windows on CA detection. AUROC values (mean ± standard deviation) found by considering features from an increasing number of time windows and including therefore progressively more seconds of CASEs are presented in **Table 7.3**. In this table a stepwise increase in the mean AUROC values for all detection models is observed, for both cross-validation methods.

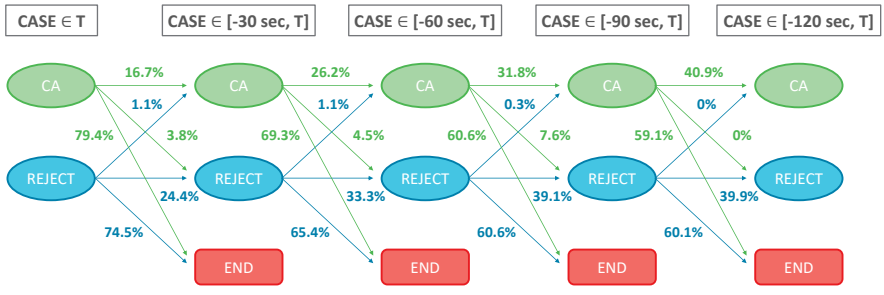


Figure 7.3: Representation of the first four stages of transitions between all central apnea-suspected events (CASEs) included in this study. Annotations returned by the clinical experts during the second round of annotation are used to create this representation. END is also added to indicate the ending of a cluster. Considering the first transition as an example, clusters starting from a CA were followed by a CA 16.7% of the time, were followed by a rejection 3.8% of the times and were not followed by another CASE in the following 30 s 79.4% of the times. Only the most likely transitions are represented (e.g., transition from CA to error is not represented due to extremely low likelihood of this transition).

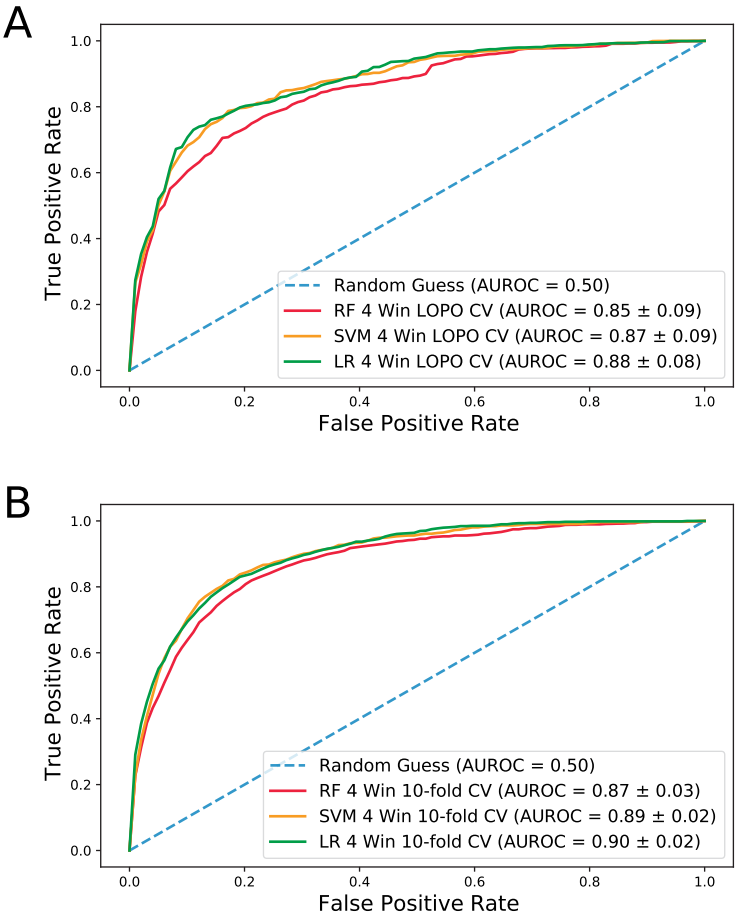


Figure 7.4: Mean receiver operating characteristic (ROC) curves computed with all detection models based on random forest (RF), support vector machines (SVM) and logistic regression (LR) with elastic net penalty using features from windows [-30 s, 0 s), [-25 s, 5 s), [-20 s, 10 s) and [-15 s, 15 s) with the leave-one-patient-out cross-validation (LOPO CV) (A) and 10-fold cross-validation (10-fold CV) (B).

As shown in **Figure 7.4A**, the best performing detection model by means of LOPO CV was identified in LR with elastic net penalty trained using features from all 4 time windows, since it showed the highest mean AUROC. Feature relevance was investigated by extracting the coefficients from the detection model. These are log odds ratios and while positive values indicate increased risk towards CA, negative values indicate increased risk towards rejection. Median log odds ratios computed considering all folds for the twenty most relevant features, ranked considering their absolute values are shown in **Figure 7.5**. In general, relevance for features extracted from the respiration

RRE was found to be generally the highest among all features included in this study. Different features extracted from the RR-intervals also proved to be relevant for the detection whereas features extracted from the SII, CRC and SpO_2 showed lower median log odds ratios. The most frequently selected set of hyperparameters consisted of C = 0.005 and L1 ratio = 1, and was selected for 16 out of 20 folds.

In the third experiment we evaluated whether the detection was performed differently for different groups of patients while using the detection model made using LR with elastic net penalty and LOPO CV. AUROC values for all the patients in the test set are shown in **Figure 7.6**, represented in blue for control patients and in red for LOS patients. No statistical significance while comparing matching LOS and control patients was found (p-value > 0.05). AUROC values for the 5 patients experiencing a low number of CAs (count of annotated CAs ≤ 5) showed high values (mean AUROC = 0.948, min AUROC = 0.895, max AUROC = 1) while a wide variety of AUROC values was found in the six patients with high number of CAs (count of CAs ≥ 50) (mean AUROC = 0.850, min AUROC = 0.659, max AUROC = 0.951).

Table 7.3: AUROC values (mean ± standard deviation) found by considering features an increasing number of consecutive time windows for each central apnea-suspected event (CASE). Results found for detection models based on logistic regression (LR) with elastic net penalty, random forest (RF) and support vector machines (SVM), as well as both cross-validation methods, leave-one-patient-out cross-validation (LOPO CV) and 10-fold cross-validation (10-fold CV), are shown. The following time windows including progressively more seconds of CASEs were considered to obtain these results: [-30 s, 0 s), [-25 s, 5 s), [-20 s, 10 s) and [-15 s, 15 s).

LOPO CV		LR	RF	SVM
1 Window	AUROC (Mean ± Std)	0.73 ± 0.12	0.70 ± 0.11	0.74 ± 0.12
2 Windows	AUROC (Mean ± Std)	0.80 ± 0.10	0.76 ± 0.11	0.79 ± 0.10
3 Windows	AUROC (Mean ± Std)	0.87 ± 0.08	0.81 ± 0.09	0.86 ± 0.08
4 Windows	AUROC (Mean ± Std)	0.88 ± 0.08	0.85 ± 0.09	0.87 ± 0.09

10-fold CV		LR	RF	SVM
1 Window	AUROC (Mean ± Std)	0.80 ± 0.02	0.78 ± 0.02	0.80 ± 0.02
2 Windows	AUROC (Mean ± Std)	0.85 ± 0.03	0.81 ± 0.03	0.84 ± 0.03
3 Windows	AUROC (Mean ± Std)	0.89 ± 0.02	0.86 ± 0.03	0.88 ± 0.03
4 Windows	AUROC (Mean ± Std)	0.90 ± 0.02	0.87 ± 0.03	0.89 ± 0.02

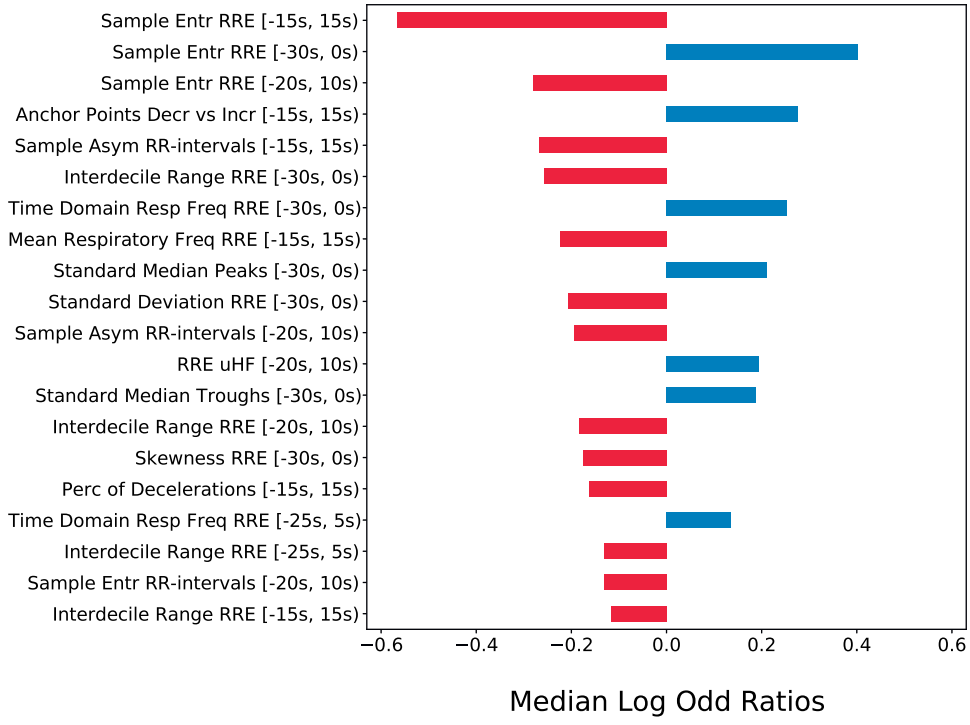


Figure 7.5: Feature relevance for the logistic regression (LR) with elastic net penalty fed with features from time windows [-30 s, 0 s), [-25 s, 5 s), [-20 s, 10 s) and [-15 s, 15 s) and trained using leave-one-patient-out cross-validation (LOPO CV). Feature relevance was investigated by extracting the coefficients, indicating the log odds ratios. Median log odds ratios were then computed and were ranked considering their absolute values to represent the twenty most relevant features.

In our fourth experiment, three different FPR (0.1, 0.2 and 0.3) were defined to extract the confusion matrices for the same detection model based on LR with elastic net penalty and LOPO CV. Confusion matrices as well as other metrics are shown in **Table 7.4**. A FPR = 0.3 in particular allowed to detect CAs with a recall (TPR) = 0.78, a precision = 0.42 and a miss rate (FNR) = 0.22. Since we considered as highest priority to avoid missing too many CAs even at the cost of a slight increase of false CA detection, we chose to use this FPR value for the subsequent and final experiment.

In our fifth experiment, percentages of correct detection for different events found with FPR = 0.3 were computed. Results are shown in **Figure 7.7** together with the total number of events considered for each group. Percentages of correct detections were found to be equal to 78.2% for all CAs and very similar values were also found for CAs followed by a desaturation (79.0%) or decrease in SpO_2 (75.7%).

Significantly higher values percentages of correct detections were found for CAs followed by a bradycardia (93.4%), by a decrease in HR (93.9%) and by both a bradycardia and a desaturation (95.2%). In addition, percentages of correct detection were found to be higher for CAs with length 20 s ≤ L < 30 s (88.3%) or for CAs with length L ≥ 60 s (91.7%). Finally, percentages of correct detection were found to be higher for CASEs originally agreed upon by clinical experts (82.9%) compared to CASEs for which there was no initial agreement (60.1%).

7.4. Discussion

Correct identification of apnea of prematurity is important as a correlation has been described between apnea of prematurity and neurodevelopmental impairment as well as other pathologies, such as sepsis, that can represent a major threat to the life of premature infants [2], [4]–[6]. Therefore, we developed a detection model based on machine learning which shows that it is feasible to automatically detect true CAs and to reduce the occurrence of high load of false apnea alarms observed in the current clinical practice.

As a result of an investigation of CASEs extracted from our population of premature infants, we found a high likelihood for CASEs to appear in clusters, as 27.8% of the total count of CASEs was found to be preceded by another CASE in the previous 30 s. This result was found to be in agreement with previous literature studies which also indicated that periodic apneas can be found in premature infants [6], [14], [20]. In addition to this we observed that consistency in annotations can be found for CASEs appearing in a cluster. In particular, we found a far higher likelihood for CAs to be clustered together rather than with other apneic events annotated as rejections. Consistency of annotations within clusters of CASEs might also be considered for the improvement of detection models. Future studies could also provide more precise annotations for rejections (e.g., separating obstructive and mixed apneas) and investigate whether consistency characterizes other clusters as well.

Furthermore, our first machine learning experiment showed very similar AUROC values for the different detections models that were implemented based on LR with elastic net penalty, RF and SVM, considering features extracted from all 4 time windows and both cross-validation methods.

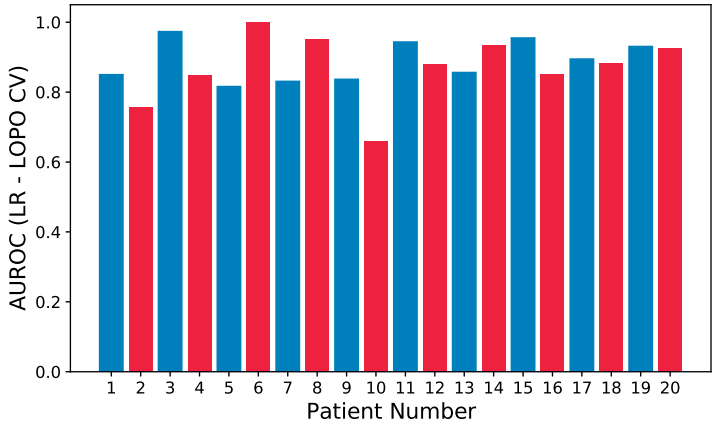


Figure 7.6: AUROC values computed for each patient in the test set with the detection model based on logistic regression (LR) with elastic net penalty using features from windows [-30 s, 0 s), [-25 s, 5 s), [-20 s, 10 s) and [-15 s, 15 s) with the leave-one-patient-out cross-validation (LOPO CV). AUROC values are represented in blue for control patients and in red for LOS patients.

Table 7.4: Confusion matrices found considering a False Positive Rate (FPR) equal to (A) 0.1, (B) 0.2 and (C) 0.3 in the mean ROC curve for logistic regression (LR) with elastic net penalty, estimated labels by this detection model using the corresponding threshold and the annotations from the clinical experts as true labels.

A FPR = 0.1		Estimated Labels	
		Central Apnea	Reject
True Labels (Annotations)	Central Apnea	529	346
	Reject	328	4130
Recall (TPR)	0.60	Precision (PPV)	0.62
Miss Rate (FNR)	0.39	Specificity (TNR)	0.93

B FPR = 0.2		Estimated Labels	
		Central Apnea	Reject
True Labels (Annotations)	Central Apnea	633	242
	Reject	662	3796
Recall (TPR)	0.72	Precision (PPV)	0.49
Miss Rate (FNR)	0.28	Specificity (TNR)	0.85

C FPR = 0.3		Estimated Labels	
		Central Apnea	Reject
True Labels (Annotations)	Central Apnea	684	191
	Reject	937	3521
Recall (TPR)	0.78	Precision (PPV)	0.42
Miss Rate (FNR)	0.22	Specificity (TNR)	0.79

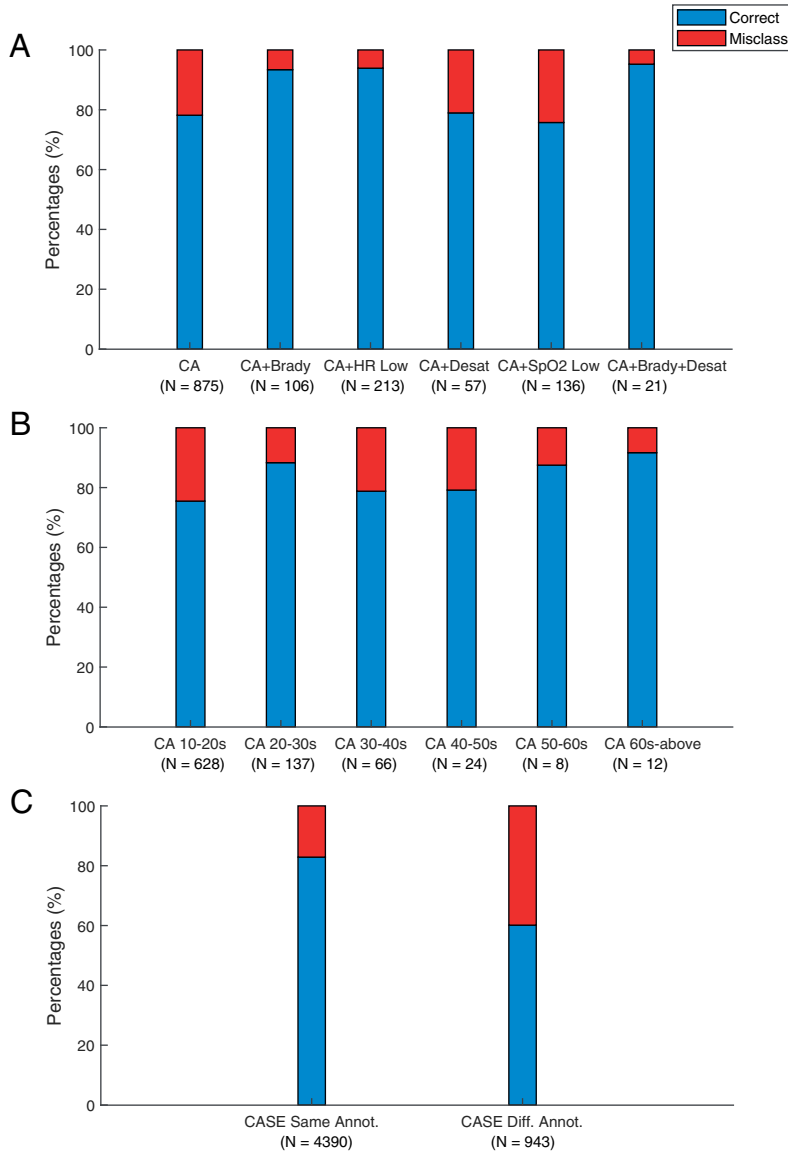


Figure 7.7: Percentages of correct detections (recall) performed by logistic regression (LR) with elastic net penalty trained using leave-one-patient-out cross-validation (LOPO CV) on CASEs present in the test set when a FPR = 0.3 was selected. In particular, in (A) all central apneas (CAs), CAs followed by a decrease in heart rate (HR) or oxygen saturation (SpO_2), in (B) CAs with different lengths and in (C) central apnea-suspected events (CASEs) agreed and disagreed upon by clinical experts during the first round of annotations are represented.

AUROC values found with RF and SVM were found to be higher than in the promising previous literature studies which made use of the same machine learning algorithms

[22], [23]. Our AUROC values found by means of LR with elastic net penalty are instead more difficult to compare to results achieved in the past. This is because, to the best of our knowledge, LR was only used in combination with a Gaussian mixture model by Zuzarte et al. with the aim of performing apnea prediction [6]. Furthermore, this previous model allowed to perform a live apnea prediction, while in our study we considered performing apnea detection based on single CASEs, similarly to what was done in previous studies who used RF and SVM [22], [23]. In this study mean AUROC values were also found to be consistent for both cross-validation methods. However, it is important to highlight the higher uncertainty, indicated by the AUROC standard deviation, found with LOPO CV, a result that is related to different count of CASEs identified per patient. On the other hand, methods based on testing the detection model on new patients, like LOPO CV, are more similar to the regular clinical practice, where a detection model based on machine learning should be trained with previously admitted patients and used (i.e tested) on a newly admitted patient. In addition to this, AUROC values found using 10-fold CV, which are due to overfitting resulting from the use of CASEs from the same patient in both the training and test set, are only slightly higher compared to those obtained with LOPO CV, giving further validity for the use of a model based on LOPO CV in the other experiments included in this study.

In the second machine learning experiment we found out that the highest AUROC values were found with features extracted from up to the first 15 s of each CASE, irrespectively of the detection model and the cross-validation method that was used. It must be noted, though, that AUROC values were quite similar by including only features up to the first 10 s of each CASE. This result is particularly significant for the clinical practice, as all definitions present in literature require a minimum of 10 s of a cessation of breathing to define apnea of prematurity [7]–[11]. Our detection model therefore not only detects apneas based on vital signs but predicts the occurrence of an upcoming true CA alarm.

The investigation of feature relevance performed for LR with elastic net penalty, features from all 4 time windows and LOPO CV showed higher median log odds ratios for features extracted from the respiration RRE compared to all the other features. This result is expected as a CA is likely diagnosed by observations of the CI, with consequences in HR and SpO_2 appearing just at a later stage. Nonetheless, features extracted from the RR-intervals also provided a significant contribution to the detection model whereas features extracted from the CRC and SpO_2 were found to be in general less relevant. In contrast to recent literature [6], in our study the motion signal (calculated in our study as SII) shows only a modest contribution, which may be explained by the fact that we focused our attention on apneas with a central origin, whereas patient movement is associated with apneas of an obstructive origin [52], and by the fact that the SII is extracted from the ECG signal, a signal on which the patient monitor applies different

methods to prevent motion artifacts [32]. Future studies could therefore investigate whether wavelet-based algorithms, used for instance by Zuzarte et al. [6], are more suitable than SII to contribute to the detection of apnea of prematurity.

In our third machine learning experiment no difference was found by comparing AUROC values for LOS versus control patients, indicating the possibility of applying LR with elastic net penalty for CA detection irrespectively of the patient group. We also noticed high AUROC values for 5 patients experiencing a low number of CAs. With an in-depth analysis focused on the interpretation of this result, we found out that CA records for patients with few CAs used as a test set generally presented high probability values, irrespectively of the machine learning model that was used. As soon as high threshold values (low FPR values) were trespassed, an increase in TPR and consequently in the AUROC values became immediately noticeable. We also speculate that this result might have a clinical reason, since we noticed a very regular breathing pattern in this group of patients. Because of this reason, a disruption in this breathing pattern could be easily recognized. Patients with several CAs often showed instead a less regular breathing pattern, which could have led to more rejections being wrongly identified as CAs and at the same time made CAs harder to be detected.

In our fourth machine learning experiment, a FPR = 0.3 was selected as an optimal threshold for our best performing detection model. This choice was performed since in the medical field it is preferred to correctly detect as many true events as possible (i.e. high recall), even at the cost of having a lower precision and triggering therefore more false alarms [9]. Recent studies have indicated a precision equal to 0.35 for the current monitoring systems available in clinical practice (i.e. up to 65% of the apnea alarms currently found in NICUs are false) [17]–[19]. Using our detection model with a FPR = 0.3 we improved the precision to 0.42, reducing the number of false alarms by 66% compared to the application of the optimized algorithm for central apnea detection only, an algorithm derived from the one originally described by Lee et al. which showed a lower number of false positives compared to what is obtained using the current monitoring techniques [17].

Percentages of correct detections (recall) were further examined in the fifth machine learning experiment and were found to be particularly high for CAs followed by a decrease in HR (80 bpm < HR ≤ 100 bpm), a bradycardia and both a bradycardia and a desaturation, which are considered the most concerning for the well-being of the infants [2], [14]. Our results are in line with the findings of Williamson et al., who showed promising results with an AUROC equal to 0.79 while applying a different machine learning algorithm (i.e., quadratic classifier) [20]. Lower percentages were instead found when CAs were followed by a decrease in SpO_2, a result that can be related to the low

relevance found for features extracted from this signal. An explanation for this result was given by Di Fiore et al., who indicated that false desaturation alarms can be reduced by using long averaging time, equal to at least 16 s [2], which is nowadays typically used in most clinical practices. Exceeding the first 15 s of CASE for feature extraction might therefore be one solution to improve detection of CAs followed by a decrease in saturation. On the other hand, since 15 s of cessation of breathing are sufficient to completely define apnea of prematurity [9]–[11], the inclusion of this requirement would have prevented us from the possibility of performing an actual prediction of CAs and was therefore not considered.

Another solution that can be explored in future studies is the definition of new features capable to estimate possible future decrease in SpO_2, including for instance features computed in the frequency domain which proved to be useful for diagnosing obstructive sleep apnea syndrome in adults [53]. Finally, it is important to note that the percentages of correct detections were lower in case the clinical experts disagreed during the first round of annotations (recall equal to 60.1% instead of 82.9%, found in case of agreement).

This study has some noteworthy limitations. First of all, a limited sample size of patients and CASEs, which was selected considering the time availability for clinicians to provide annotations in different rounds, as this annotation process – even though guided by a user interface – was still a very time-consuming task for the clinical experts. This was also the reason why annotations were made by no more than two clinical experts. Second, in this study detection models were only used retrospectively. Also, the two-step approach, with first detection of cessation of breathing and then CA is slow for implementation. Using a full dataset including all 30 s long time windows, independent of the presence of a CASE, would be of interest. However, this would result in a highly unbalanced dataset. Third, the inclusion of more refined features from different signals, including the SpO_2, and the inclusion of other features to quantify body motion as well as the gestational age, postnatal days and birth weight could also improve the detection. Fourth, validation on data coming from a different dataset or another hospital was not investigated in this study. Since there is currently no public dataset available that includes annotations for CAs in premature infants as well as all the signals included in this study, future studies could include data from a different dataset to provide a further indication on the validity of the current results. However, we consider our study to be valid as a first feasibility study to verify our methodology before a larger-scale study including more data from more patients and different datasets. Finally, only machine learning models characterized by an easier transfer of knowledge regarding the results and that could cover regression, clustering and tree-based algorithms were included in this study. Future studies, more focused on the technical aspects, could however be

considered to expand on the selected models for instance by including deep neural networks.

7.5. Conclusion

In this study different detection models based on machine learning were applied in order to detect CAs among CASEs. We showed that CAs can be automatically detected correctly while reducing the number of false alarms, using a detection model based on well-known machine learning algorithms. A threshold for the FPR in the mean receiver-operating-characteristic curve equal to 0.3 led to a percentage of 78.2% correct detections for all CAs and 93.4% and 95.2% for the most critical and clinically relevant CAs, namely those followed by bradycardia or bradycardia and desaturation.

References

[1] E. C. Eichenwald and Committee on Fetus and Newborn, "Apnea of Prematurity," *Pediatrics*, vol. 137, no. 1, 2016.

[2] J. M. Di Fiore, R. J. Martin, and E. B. Gauda, "Apnea of prematurity - Perfect storm," *Respir. Physiol. Neurobiol.*, vol. 189, no. 2, pp. 213–222, 2013.

[3] B. A. Sullivan and K. D. Fairchild, "Vital signs as physiomarkers of neonatal sepsis," *Pediatr. Res.*, vol. 91, no. 2, pp. 273–282, 2022.

[4] S. M. Gephart, J. M. McGrath, J. A. Effken, and M. D. Halpern, "Necrotizing Enterocolitis Risk," *Adv. Neonatal Care*, vol. 12, no. 2, pp. 77–87, 2012.

[5] A. Janvier, M. Khairy, A. Kokkotis, C. Cormier, D. Messmer, and K. J. Barrington, "Apnea is associated with neurodevelopmental impairment in very low birth weight infants," *J. Perinatol.*, vol. 24, no. 12, pp. 763–768, 2004.

[6] I. Zuzarte, D. Sternad, and D. Paydarfar, "Predicting apneic events in preterm infants using cardio-respiratory and movement features," *Comput. Methods Programs Biomed.*, vol. 209, p. 106321, 2021.

[7] N. N. Finer, R. Higgins, J. Kattwinkel, and R. J. Martin, "Summary proceedings from the apnea-of-prematurity group," *Pediatrics*, vol. 117 (Suppl, no. 3, pp. S47–S51, 2006.

[8] J. Zhao, F. Gonzalez, and D. Mu, "Apnea of prematurity: From cause to treatment," *Eur. J. Pediatr.*, vol. 170, no. 9, pp. 1097–1105, 2011.

[9] S. B. Amin and B. Erica, "Monitoring Apnea of Prematurity: Validity of Nursing Documentation and Bedside Cardiorespiratory Monitor," *Am J Perinatol*, vol. 30, no. 8, pp. 643–648, 2013.

[10] D. E. Elder, A. J. Campbell, and D. Galletly, "Current definitions for neonatal apnoea: Are they evidence based?," *J. Paediatr. Child Health*, vol. 49, no. 9, 2013.

[11] N. N. Finer, K. J. Barrington, B. J. Hayes, and A. Hugh, "Obstructive, mixed, and central apnea in the neonate: Physiologic correlates," *J. Pediatr.*, vol. 121, no. 6, pp. 943–950, 1992.

[12] S. Picone, R. Aufieri, and P. Paolillo, "Apnea of prematurity: challenges and solutions," *Res. Reports Neonatol.*, p. 101, 2014.

[13] O. P. Mathew, "Apnea of prematurity: Pathogenesis and management strategies," *J. Perinatol.*, vol. 31, no. 5, pp. 302–310, 2011.

[14] C. F. Poets, "Apnea of prematurity: What can observational studies tell us about pathophysiology?," *Sleep Med.*, vol. 11, no. 7, pp. 701–707, 2010.

[15] T. Carbone, L. C. Marrero, J. Weiss, M. Hiatt, and T. Hegyi, "Heart Rate and Oxygen Saturation Correlates of Infant Apnea," *J. Perinatol.*, vol. 19 (1), pp. 44–47, 1999.

[16] D. E. Weese-Mayer, R. T. Brouillette, A. S. Morrow, L. P. Conway, L. M. Klemka-Walden, and C. E. Hunt, "Assessing validity of infant monitor alarms with event recording," *J. Pediatr.*, vol. 115, no. 5 PART 1, pp. 702–708, 1989.

[17] H. Lee *et al.*, "A new algorithm for detecting central apnea in neonates," *Physiol. Meas.*, vol. 33, no. 1, pp. 1–17, 2012.

[18] B. D. Vergales *et al.*, "Accurate automated apnea analysis in preterm infants," *Am. J. Perinatol.*, vol. 31, no. 2, pp. 157–162, 2014.

[19] M. A. Mohr *et al.*, "Very long apnea events in preterm infants," *J. Appl. Physiol.*, vol. 118, no. 5, pp. 558–568, 2015.

[20] J. R. Williamson, D. W. Bliss, D. W. Browne, P. Indic, E. Bloch-Salisbury, and D. Paydarfar, "Using physiological signals to predict apnea in preterm infants," *Conf. Rec. - Asilomar Conf. Signals, Syst. Comput.*, pp. 1098–1102, 2011.

[21] S. Y. Belal, A. J. Emmerson, and P. C. W. Beatty, "Automatic detection of apnoea of prematurity," *Physiol. Meas.*, vol. 32, no. 5, pp. 523–542, 2011.

[22] R. D. Shirwaikar, U. D. Acharya, K. Makkithaya, M. Surulivelrajan, and L. E. S. Lewis, "Machine learning techniques for neonatal apnea prediction," *J. Artif. Intell.*, vol. 9, no. 1–3, pp. 33–38, 2016.

[23] N. Mago, S. Srivastava, R. D. Shirwaikar, U. D. Acharya, L. E. S. Lewis, and M. Shivakumar, "Prediction of Apnea of Prematurity in neonates using Support Vector Machines and Random Forests," *Proc. 2016 2nd Int. Conf. Contemp. Comput. Informatics, IC3I 2016*, pp. 693–697, 2016.

[24] K. Lim, H. Jiang, A. P. Marshall, B. Salmon, T. J. Gale, and P. A. Dargaville, "Predicting Apnoeic Events in Preterm Infants," *Front. Pediatr.*, vol. 8, no. September, pp. 1–7, 2020.

[25] R. D. Shirwaikar, D. Acharya U, K. Makkithaya, S. M, S. Srivastava, and L. E. S. Lewis U, "Optimizing neural networks for medical data sets: A case study on neonatal apnea prediction," *Artif. Intell. Med.*, vol. 98, no. July, pp. 59–76, 2019.

[26] G. Varisco *et al.*, "Optimized Detection of Central Apneas Preceding Late-Onset Sepsis in Premature Infants," *EMBC - Curr. under Rev.*, 2021.

[27] L. Cabrera-Quiros *et al.*, "Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning," *Crit. Care Explor.*, vol. 3, no. 1, p. e0302, 2021.

[28] E. M. Edwards, D. E. Y. Ehret, R. F. Soll, and J. D. Horbar, "Vermont Oxford network: A worldwide learning community," *Transl. Pediatr.*, vol. 8, no. 3, pp. 182–192, 2019.

[29] M. P. Griffin and J. R. Moorman, "Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis," *Pediatrics*, vol. 107, no. 1, pp. 97–104, 2001.

[30] K. Fairchild *et al.*, "Clinical associations of immature breathing in preterm infants: Part 1-central apnea," *Pediatr. Res.*, vol. 80, no. 1, pp. 21–27, 2016.

[31] J. R. Williamson, D. W. Bliss, D. W. Browne, P. Indic, E. Bloch-Salisbury, and D. Paydarfar, "Individualized apnea prediction in preterm infants using cardio-respiratory and movement signals," *2013 IEEE Int. Conf. Body Sens. Networks, BSN 2013*, pp. 5–10, 2013.

[32] R. Joshi, D. Kommers, L. Oosterwijk, L. Feijs, C. Van Pul, and P. Andriessen, "Predicting Neonatal Sepsis Using Features of Heart Rate Variability, Respiratory Characteristics, and ECG-Derived Estimates of Infant Motion," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 3, pp. 681–692, 2020.

[33] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, 2006.

[34] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, "Heart rate variability Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, no. 17, pp. 354–381, 1996.

[35] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Front. Public Heal.*, vol. 5, no. September, pp. 1–17, 2017.

[36] B. P. Kovatchev, L. S. Farhy, H. Cao, M. P. Griffin, D. E. Lake, and J. R. Moorman, "Sample Asymmetry Analysis of Heart Rate Characteristics with Application to Neonatal Sepsis and Systemic Inflammatory Response Syndrome," *Pediatr. Res.*, vol. 54, no. 6, pp. 892–898, 2003.

[37] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy maturity in premature infants Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. Physiol. Hear. Circ. Physiol.*, vol. 278, pp. H2039–H2049, 2000.

[38] C. C. Mayer, M. Bachler, M. Hörtenhuber, C. Stocker, A. Holzinger, and S. Wassertheurer, "Selection of entropy-measure parameters for knowledge discovery in heart rate variability data," *BMC Bioinformatics*, vol. 15, no. 6, pp. 1–11, 2014.

[39] D. R. Kommers *et al.*, "Features of Heart Rate Variability Capture Regulatory Changes During Kangaroo Care in Preterm Infants," *J. Pediatr.*, vol. 182, pp. 92-98.e1, 2017.

[40] A. Bauer *et al.*, "Phase-rectified signal averaging detects quasi-periodicities in non-stationary data," *Phys. A Stat. Mech. its Appl.*, vol. 364, pp. 423–434, 2006.

[41] J. W. Kantelhardt *et al.*, "Phase-rectified signal averaging for the detection of quasi-periodicities and the prediction of cardiovascular risk and the prediction of cardiovascular risk," *CHAOS*, vol. 17, no. 015112, pp. 1–9, 2007.

[42] R. Joshi *et al.*, "Cardiorespiratory coupling in preterm infants," *J. Appl. Physiol.*, vol. 126, no. 1, pp. 202–213, 2019.

[43] P. Indic, E. B. Salisbury, D. Paydarfar, E. N. Brown, and R. Barbieri, "Interaction between heart rate variability and respiration in preterm infants," *Comput. Cardiol.*, vol. 35, pp. 57–60, 2008.

[44] E. Longin, T. Gerstner, T. Schaible, T. Lenz, and S. König, "Maturation of the autonomic nervous system: Differences in heart rate variability in premature vs. term infants," *J. Perinat. Med.*, vol. 34, no. 4, pp. 303–308, 2006.

[45] Z. Peng *et al.*, "Body Motion Detection in Neonates Based on Motion Artifacts in Physiological Signals from a Clinical Patient Monitor," in *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2021.

[46] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, "Analyzing respiratory effort amplitude for automated sleep stage classification," *Biomed. Signal Process. Control*, vol. 14, no. 1, pp. 197–205, 2014.

[47] X. Long, P. Fonseca, R. M. Aarts, R. Haakma, and J. Foussier, "Modeling cardiorespiratory interaction during human sleep with complex networks," *Appl. Phys. Lett.*, vol. 105, no. 20, 2014.

[48] M. M. Kabir *et al.*, "Cardiorespiratory phase-coupling is reduced in patients with obstructive sleep apnea," *PLoS One*, vol. 5, no. 5, 2010.

[49] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuño, "From time series to complex networks: The visibility graph," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 13, pp. 4972–4975, 2008.

[50] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *J. R. Stat. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.

[51] G. Varisco *et al.*, "Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU," *Acta Paediatr. Int. J. Paediatr.*, vol. 110, no. 4, pp. 1141–1150, 2021.

[52] E. S. Katz, R. B. Mitchell, and C. M. D'Ambrosio, "Obstructive sleep apnea in infants," *Am. J. Respir. Crit. Care Med.*, vol. 185, no. 8, pp. 805–816, 2012.

[53] D. Álvarez, R. Hornero, J. Víctor Marcos, and F. Delcampo, "Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 12, pp. 2816–2824, 2010.

Appendix 7.1: List of extracted features

List of all the extracted features from the RR-intervals, Signal Instability Index (SII), respiration Ribcage Respiratory Effort (RRE), Cardiorespiratory Coupling (CRC) and SpO_2 signals.

Features extracted from the RR-intervals

Feature	Description
Mean RR	Mean RR-interval [34]
SDNN	Standard deviation of the RR-intervals [34], [35]
RMSSD	Root mean square of successive differences between adjacent normal RR-intervals [34], [35]
Sample Asymmetry of the RR-intervals	Measure of the asymmetry in the histogram of RR-intervals [36]
Sample Entropy of the RR-Intervals	Sample entropy of the RR-intervals, computed with m = 2 and r = 0.2 · o [37], [38]
Percentage of Decelerations	Percentage of the RR-intervals longer than the mean RR-interval from the current time window [27], [39]
Standard Deviation of Decelerations	Standard deviation of all RR-intervals that contribute to the Percentage of Decelerations [39]
Immediate Deceleration Response (IDR)	Difference between the maximum and minimum values of RR-intervals in the 5 data points after (including the anchor points) and before the anchor point from the PRSA curve computed considering increase events (T=1) [40], [42]
Immediate Acceleration Response (IAR)	Difference between the maximum and minimum values of RR-intervals in the 5 data points after (including the anchor points) and before the anchor point from the PRSA curve computed considering decrease events (T=1) [40], [42]
Average Deceleration Response (ADR)	Difference between the mean value of the 10 RR-intervals after (including the anchor point) and the 10 RR-intervals before the anchor point from the PRSA curve computed considering increase events (T=1) [40], [42]
Average Acceleration Response (AAR)	Difference between the mean value of the 10 RR-intervals after (including the anchor point) and the 10 RR-intervals before the anchor point from the PRSA curve computed considering decrease events (T=1) [40], [42]
Slope of the Deceleration Response (SDR)	Slope of the line joining the maximum and minimum values of the RR-intervals corresponding to the IDR computed from PRSA curve computed considering increase events (T=1) [40], [42]

Slope of the Acceleration Response (SAR)	Slope of the line joining the maximum and minimum values of the RR-intervals corresponding to the IAR computed from PRSA curve computed considering decrease events (T=1) [40], [42]
Anchor Points Decrease versus Increase	Feature computed by dividing the difference between the count of anchor points from the PRSA with decrease events and with increasing events by the count of RR-intervals and then multiplying this value by 100

Features extracted from the RR-intervals (resampled at 250 Hz)

Feature	Description
RR-intervals LF	PSD computed from the FFT considering the low frequency range (0.01-0.15 Hz) [43]
RR-intervals HF	PSD computed from the FFT considering the high frequency range (0.15-0.45 Hz) [43]
RR-intervals sHF	PSD computed from the FFT considering the super high frequency range (0.45-0.7 Hz) [43]
RR-intervals uHF	PSD computed from the FFT considering the ultra high frequency range (0.7-1.5 Hz) [43]
RR-intervals LF/HF Ratio	Ratio between the values for the PSD computed from the FFT in the LF and HF ranges [33], [43]
Mean Respiratory Frequency RR-intervals	Frequency of maximum power in the range 0.15-0.45 Hz [33]

Features extracted from the Signal Instability Index (SII)

Feature	Description
Mean SII	Mean value of the SII
Standard Deviation of the SII	Standard deviation of the SII [27], [32]
Skewness of the SII	Skewness of the SII [27], [32]
Interdecile Range of the SII	Interdecile range of the SII [27], [32]

Features extracted from the respiration Ribcage Respiratory Effort (RRE)

Feature	Description
Mean RRE	Mean value of the respiration RRE
Standard Deviation of the RRE	Standard deviation of the respiration RRE
Skewness of the RRE	Skewness of the respiration RRE
Interdecile Range of the RRE	Interdecile range of the respiration RRE
Sample Entropy of the RRE	Sample entropy of the respiration RRE, computed with m = 2 and r = 0.2 · o [37], [38]
Envelope Power	Computed by taking the standard deviation of the peaks for an epoch and of the standard deviation of the troughs. The mean value is computed and divided by the standard deviation of the respiration RRE for the epoch [33]
Breath-by-Breath Correlation	Computed by defining as breath cycle the time from the trough of one breath to the trough of the next. After computing the maximum cross-correlation value for adjacent breaths, this value is divided by the maximum of the energy of either breath alone to normalize the maximum cross-correlation value. The maximum cross-correlation values, for all pairs of adjacent breaths, are then averaged [33]
Breath Length Variation	Computed considering the standard deviation of the time between peak locations and the standard deviation of the time between trough locations and calculating the mean of these two values [33]
Time Domain Respiratory Frequency	Inverse of the mean time between adjacent peaks and between adjacent troughs [33]
Standardized Median of the Peaks	Computed by dividing the median of the peaks in the respiration RRE by their interquartile range [46]
Standardized Median of the Troughs	Computed by dividing the median of the troughs in the respiration RRE by their interquartile range [46]
RRE LF	PSD computed from the FFT considering the low frequency range (0.01-0.15 Hz) [43]
RRE HF	PSD computed from the FFT considering the high frequency range (0.15-0.45 Hz) [43]
RRE vHF	PSD computed from the FFT considering the super high frequency range (0.45-0.7 Hz) [43]
RRE uHF	PSD computed from the FFT considering the ultra high frequency range (0.7-1.5 Hz) [43]
RRE LF/HF Ratio	Ratio between the values for the PSD computed from the FFT in the LF and HF ranges [33], [43]
Mean Respiratory Frequency RRE	Frequency of maximum power in the range 0.15-0.45 Hz [33]

Features extracted from the Cardiorespiratory Coupling (CRC)

Feature	Description
Mean Degree of the Nodes	Mean node degree computed from the Nonlinear Visibility Graph by averaging the degrees over the nodes with a 3.5 min window, with the last 30 s overlapping with the window used for the detection [47], [49]
Degree Variation of the Nodes	Standard deviation of the node degrees computed from the Nonlinear Visibility Graph [47], [49]
Assortativity Coefficient	Computed with the following equation: $\xi = \frac{M^{-1}\sum_i \alpha_i\beta_i - \left[M^{-1}\sum_i \frac{1}{2}(\alpha_i+\beta_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(\alpha_i^2+\beta_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(\alpha_i+\beta_i)\right]^2}$ considering a network including a total of M edges, the i-th edge connects two nodes with degree of α_i and β_i at their ends [47], [49]

Features extracted from the SpO_2

Feature	Description
Mean SpO_2	Mean value of the SpO_2
Standard Deviation of the SpO_2	Standard deviation of the SpO_2
Slope of the SpO_2	Slope of the linear fitting of all the SpO_2 points

CHAPTER 8

Multichannel data-driven central apnea detection using machine learning

Abstract

Apnea of prematurity is a developmental disorder that occur in premature infants. Monitoring of this condition, which is performed in neonatal intensive care units (NICUs) by detecting central apneas (CAs) in the respiratory traces, is characterized by a high number of false alarms. In order to improve the detection of CAs, we have previously developed a two-step model which performs first a detection of apneic events in the chest impedance signal by means of an optimized algorithm and then determine whether these are CAs using a detection model based on machine learning. This solution is however slow for implementation since the optimized algorithm first requires at least several seconds of a cessation of breathing to detect an apneic event. In this study we therefore investigated the feasibility of performing CA detection directly by means of a multichannel detection model based on machine learning, which would allow for a faster CA detection more suitable for the clinical practice.

From a dataset consisting of 48 h of data from 10 premature infants we extracted 47 features from physiological signals derived from ECG, chest impedance and oxygen saturation, using 30 s moving windows and an overlap of 5 s, creating 4-windows-sets (4 consecutive windows) for further processing. An undersampling method from literature and one undersampling method that we developed to consider distances within majority class samples, were used in order to reduce the imbalance in the dataset and returned two datasets to be used in this study. A detection model based on logistic regression with elastic net penalty and leave-patient-out cross-validation was then used with both datasets.

Mean AUROC values above 0.80 were found considering both datasets. By selecting the combination of model and dataset characterized by the highest mean AUROC and a false positive rate (FPR) threshold equal to 0.2 we were able to achieve an increase in precision (i.e., reduction of false alarms) compared to the use of the optimized algorithm included in the two-step approach without severely affecting its recall. The possibility of performing a smoothing option, by considering as true only CA annotations returned by the detection model for 3 consecutive 4-windows-sets, allowed to correctly detect a median of 88.3% of CAs per patient and a median of 74.2% of CAs within the first 25 s from their onset.

This study showed that CA detection in premature infants by means of a multichannel detection model based on machine learning is feasible in a time efficient way without the need of any preprocessing, a characteristic which allows for an implementation on clinical patient monitors.

8.1. Introduction

Neonatal intensive care units (NICUs) are the designated hospital environments where care for premature infants is provided [1]. Here, vital signs from these fragile patients are continuously monitored by monitoring devices which return alarms to warn nurses and clinicians in case of patient deterioration. This strategy allows to administer an appropriate treatment for each condition that is developed by the premature infant and to prevent further complications. Alarms are often separated into technical alarms, which are for instance used in case of loose sensors or to indicate a malfunctioning of the monitoring devices, and alarms related to vital signs, which warn nurses and clinicians when a vital sign has breached the designated optimal target range and reached an alerting or critical range. It has however been estimated that up to 80% of the alarms sounding in NICUs are false or clinically irrelevant [2], [3]. Therefore, these alarms don't provide any additional information to nurses and clinicians but instead they only increase the alarm burden in these hospital environments. This problem can also result in alarm fatigue, a condition in which nurses and clinicians can become desensitized to the alarms, with the potential risk of delayed and missed responses to critical alarms [4]–[6].

Different solutions have been proposed in the recent years to solve the problem of the high alarm burden present in NICUs. These are focused on two main strategies. The first relates to the optimization of the alarm parameters, which can for instance be customized depending on different patient profiles and of the workflow within a NICU [7]–[10]. The second strategy is instead focused on the development of new solutions and algorithms to improve the detection of important conditions that are currently associated with a high number of false alarms [11]–[13].

A condition which is characterized by a high number of false alarms in clinical practice (up to even 65%) is apnea of prematurity, a developmental disorder which occurs in premature infants and that relates to an immaturity of respiratory control [14]–[16]. Apnea of prematurity is diagnosed when several apneic events, characterized by cessations of breathing longer than 20 s or longer than 10 s in case these are followed by a bradycardia or desaturation, are detected in the respiratory traces of a premature infants [14], [17]. Apneic events are classified into central apneas (CAs), characterized by a cessation of the respiratory drive with consequent absence of effort to breath in the infant, obstructive apneas, characterized by an airway obstruction and therefore an absence of respiratory flow, as well as mixed apneas, which shared characteristics of both central and obstructive apneas [14], [18].

Different solutions to reduce the number of false apnea alarms have been proposed in the recent years. A noticeable one is the solution proposed by Lee et al., who developed an algorithm for CA detection which filters out the cardiac artifact from the chest impedance (CI) signal before computing an apnea probability function [11], [18], [19]. It has also been shown that this algorithm is characterized by a higher precision compared to the algorithms used by patient monitors in the clinical practice, which can be further optimized without affecting the high recall that characterizes this algorithm [11], [20]. Other studies made instead use of machine learning and deep learning algorithms and showed very promising results [21]–[26]. Skirwaikar et at. [21]–[23] used various supervised machine learning methods on a dataset containing demographic information, maternal covariates together with physiological values to detect the occurrence of apnea of prematurity. Zuzarte et al. [25] managed to predict apneic events accompanied by bradycardia and hypoxia, using cardiorespiratory and movement features derived from physiological waveform data extracted from 10 premature infants in a model that combined an unsupervised method (Gaussian Mixture Models) with a supervised one (Logistic Regression, LR). In addition to these, in our previous study 20 patients and 48 hours of data per patient were included to detect the occurrence of CAs using LR, random forest and support vector machines (SVM) by means of a two-step model (first, a detection of apneic events, performed by means of an optimized version of the Lee's algorithm for CA detection, followed by the use of a detection model to determine whether these were CAs or not) [26].

The aim of this new study is to verify whether it is feasible to perform CA detection by means of a multichannel detection model based on machine learning, to make it faster for the use on bedside patient monitors. The use of the optimized version of the Lee's algorithm for CA detection returns in fact apneic events only after (1) the extraction of a normalized filtered CI, (2) the computation of a continuous apnea probability function and (3) the use of several criteria to combine and discard apneic events that have already occurred based on their length and distance from other apneic events [11], [26]. Whether successful, the multichannel detection model included in this study would not require the use of the use of two-step approach for CA detection and could possibly avoid introducing additional time delays.

8.2. Methods

8.2.1. Dataset and annotations

In this study we used the same dataset as in our previous study [26] which consisted of 10 premature infants who developed late-onset sepsis (LOS) and 10 matched controls,

all characterized by a gestational age (GA) ≤ 30 weeks and admitted to the NICU of Máxima Medical Center in Veldhoven, the Netherlands, from July 2016 to December 2018 [26]. While for LOS patients we extracted the 48 h of data immediately preceding the CRASH-moment (blood Culture collection, Resuscitation, and Antibiotics Started Here), as defined by Griffin et and Moorman [27], for the matched controls we extracted the 48 h immediately preceding an equivalent-in-time CRASH-moment, which was calculated after matching the GA and postmenstrual age (PMA). Since our past study as well as the current one had a retrospective and non-invasive nature, a waiver was provided by the medical ethical committee in accordance with the Dutch law on medical research with humans (WMO).

Vital signs for each patient, measured using a Philips IntelliVue MX800 patient monitors (Philips Medical Systems, Böblingen, Germany) were obtained from a data warehouse (PIIC-iX, Data Warehouse Connect; Philips Medical Systems, Andover, MA). Three different vital signs were considered to perform apnea detection: the ECG and chest impedance (CI), which were respectively measured at 250 Hz and 62.5 Hz using three ECG leads, and the SpO_2, which was measured at 1 Hz by means of a photoplethysmogram.

Annotation of apneic events was performed in our previous study (on the same dataset) and required two main steps: 1) a first detection of apneic events, consisting of both true and false CAs (i.e., false alarms), by means of the optimized algorithm for CA detection and 2) manual annotations for the apneic events, which were annotated into CAs, Rejections (i.e., cessations in the CI which were not annotated as CAs and thus considered false alarms) and Errors (used in case of missing or corrupt signal) by two clinical experts in two rounds of annotations until a final consensus was reached [26].

8.2.2. Signal processing and feature extraction

Following the strategy used in our previous chapter, we extracted four different signals from the patients' vital signs [26]. From the ECG we extracted the RR-intervals, a measure of the regulation performed by the autonomic nervous system on the cardiovascular system which was computed by detecting the R-peaks [12], [28], [29]. From the ECG we also extracted the signal instability index (SII), a measure of patient motion which can be derived by applying a band-pass filter (0.001–0.40 Hz) to 10 s long ECG windows followed by a computation of a kernel density estimate [12], [29]. From the CI we extracted the respiration Ribcage Respiratory Effort (RRE), a filtered normalized respiration signal which was computed by subtracting the mean from the CI, applying a tenth-order Butterworth low-pass filter (cutoff equal to 0.8 Hz) to remove the HF noise, and performing an amplitude normalization using the median peak-to-trough amplitude to remove the tidal volume for each patient [30]. Finally, the cardiorespiratory coupling signal (CRC), a signal able to provide information regarding the interaction

between the cardiovascular and the respiratory system, was computed by resampling the respiration RRE in correspondence of the R-peaks in the ECG [31]. These four signals, together with the SpO_2 were then used to perform the feature extraction.

Feature extraction was performed using 30 s moving windows with an overlap of 5 s for the full dataset of premature infants [26], in order to be able to capture quick changes happening in the vital signs. A total of 47 features were computed for each moving window. In total, 20 features were extracted from the RR-intervals: these included features for the heart rate variability computed in the time-domain (e.g., Mean RR, standard deviation of the normal RR intervals), frequency-domain (e.g., power spectral densities (PSD) computed for 4 frequency ranges) as well as features derived using the phase-rectified signal averaging algorithm (e.g., immediate deceleration and acceleration response) [28], [32]–[34]. 4 features were extracted from the SII (e.g., Mean SII, Standard Deviation of the SII) [12], [29]. 17 features were extracted from the respiration RRE: these included features computed in the time-domain (e.g., Mean RRE, Skewness RRE, Breath-by-Breath Correlation) and in the frequency-domain (computed following the same frequency ranges defined for the RR-intervals) [30], [32], [35]. 3 features were extracted from the CRC: these were found after computing the Nonlinear Visibility Graph, a method able to describe a time series based on specific geometric criteria (e.g., Mean Degree of the Nodes, Degree Variation of the Nodes) [31]. Finally, 3 features were extracted from the SpO_2 (e.g., Mean SpO_2, Slope of the SpO_2). Further details on the computation of these features can be found in [26] and in **Appendix 7.1**.

8.2.3. Windows selection and undersampling

Previous studies have indicated that machine learning algorithms are sensitive to imbalance in the dataset, for instance when the minority class samples are severely outnumbered by majority class samples [19]. In small datasets, this type of imbalance can severely affect the learning process of various machine learning algorithms, which learn too much from the majority class and not enough from the minority one, as well as the evaluation of different metrics (e.g., accuracy, precision and recall) [36]. In addition to this it is important to include a sufficiently high number of minority class samples in a dataset for a machine learning algorithm to learn from these samples [36]. To generate a more balanced dataset considering the low count of CAs included in our dataset, we therefore proposed the following steps which are also described as steps in the pipeline in **Figure 8.1**.

In step 1, we excluded patients with a very low count of CAs as these patients could have otherwise had a large effect on the imbalance. Since 5 out of the 20 patients had a very low count of CAs (CAs $\leq$ 5) resulting in a maximum of 68 CA-windows-sets, these were excluded. Since we used a matched dataset based on GA and PMA in order to

exclude differences in heart rate variability features due to maturation [37], we had to exclude the matched patients as well. The dataset after the exclusion consisted of the 10 remaining patients. The characteristics of this dataset are presented in **Table 8.1**.

In step 2, from the 48 h of data from all patients a total of 345570 sets of 4 consecutive moving windows (4-windows-sets) were defined. For each moving window the 47 features defined in our previous study [26] as well as in **Appendix 7.1** were computed, resulting in a total of 188 features for each 4-windows-set.

In step 3, since all apneic events were annotated by clinical experts, we used this information to determine a label for each 4-windows-set. All 4-windows-sets that presented at least one window with a minimum overlap (> 0s) with a CA annotation were annotated as CA-windows-sets, as shown in **Figure 8.2**, resulting in 10273 CA-windows-sets in total (median count of 707 CA-windows-sets per patient, maximum count equal to 3793, minimum count equal to 161). This same procedure was applied to Error annotations resulting in Error-windows-sets which were discarded from the analysis. All remaining 4-windows-sets, even in the event of an overlap with a Rejection, were annotated as stable-windows-sets (i.e., 4-windows-sets referring to stable periods for the premature infant).

In step 4, after noticing the high imbalance that still characterized our dataset (10273 CA-windows-sets on a total of 345570 4-windows-sets), we decided to perform a further sub-selection of 4-windows-sets useful to perform CA detection. Together with all the CA-windows-sets (i.e., minority class) we decided to include **at least** the stable-windows-sets that respected one of the following conditions:

- being one of the 7 stable-windows-sets preceding the start of a CA or following its end (i.e., a stable-windows-set whose start and end were fully enclosed in the 1 min preceding or following a CA);
- being a stable-windows-set originally annotated as Rejection by the clinical experts (i.e., a stable-windows-set overlapping with an apneic event not annotated as a CA);
- being one of the 7 stable-windows-sets preceding the start of a Rejection or following its end.

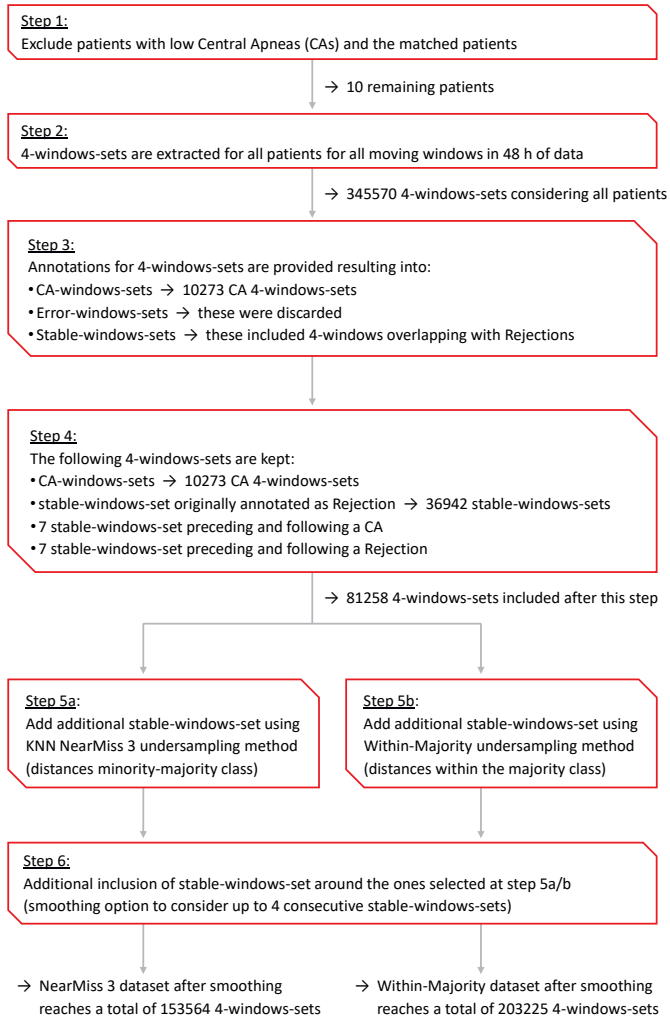


Figure 8.1: Pipeline to perform a selection of 4-windows-sets in order to generate a more balanced dataset between CA-windows-sets and stable-windows-sets.

Their inclusion was considered relevant to verify the effect of the detection around different apneic events, assuming that if a machine learning model is able to distinguish stable-windows-sets located very close or with similar characteristics to CA-windows-sets, it could also possibly perform well in detecting stable-windows-sets located very far from a CA. After this step, a total of 81258 4 windows-sets were included and the ratio between CA-windows-sets and stable-windows-sets was therefore found to be equal to around 1:7. However, this dataset was not considered to be representative enough of the original dataset since it was still quite small and did not include stable-windows-sets located very far from a CA (real stable-windows-sets).

Table 8.1: Characteristics of patients included in the analysis.

Dataset	LOS	Control	All
Count of patients	5	5	10
Gestational age (weeks) (mean ± std)	27.6 ± 1.1	27.6 ± 1.2	27.6 ± 1.1
Postnatal days (mean ± std)	13.5 ± 10.8	13.5 ± 11.7	13.5 ± 10.6
Postmenstrual age (weeks) (mean ± std)	29.6 ± 1.3	29.6 ± 1.3	29.6 ± 1.2
Birth weight (g) (mean ± std)	927.0 ± 277.9	1029.0 ± 179.9	978.0 ± 227.1
Hours per patient	48	48	48
Count of apneic events (median [IQR])	581 [387.5 – 672]	356 [243 – 465]	423 [356 – 581]
Count of CAs (median [IQR])	132 [72.7 – 210]	38 [23 – 60]	60 [36 – 132]
Count of CA-windows sets (median [IQR])	1478 [766 – 2297]	439 [273 – 679]	707 [405 – 1478]

Therefore, in step 5, different solutions to include real stable-windows-sets from the pool of remaining stable-windows-sets without building a severely unbalanced dataset were discussed. Literature regarding solutions to this problem is extensive and encompasses undersampling methods (e.g., random and informed undersampling) as well as oversampling methods (e.g., SMOTE) [36]. However, since there are no previous reports regarding the use of oversampling methods considering the signals and features included in this study, we preferred not to create additional artificial data for the minority class. We therefore added additional stable-windows-sets for each patient by means of two undersampling methods based on the K-nearest neighbor (KNN), which were used independently from each other and resulted therefore in two different datasets:

a. a dataset created with the undersampling method call NearMiss 3 by Zhang et al. [38], which selects the majority class examples showing the minimum distance to each minority class example. This was selected since we prioritized having a higher precision compared to a higher recall in order to reduce the number of false CA alarms, a characteristic that has been achieved by means of this undersampling method [38]. NearMiss 3 was implemented for each patient independently from the others. The number of majority class examples to be selected for each minority class example was set to 5;

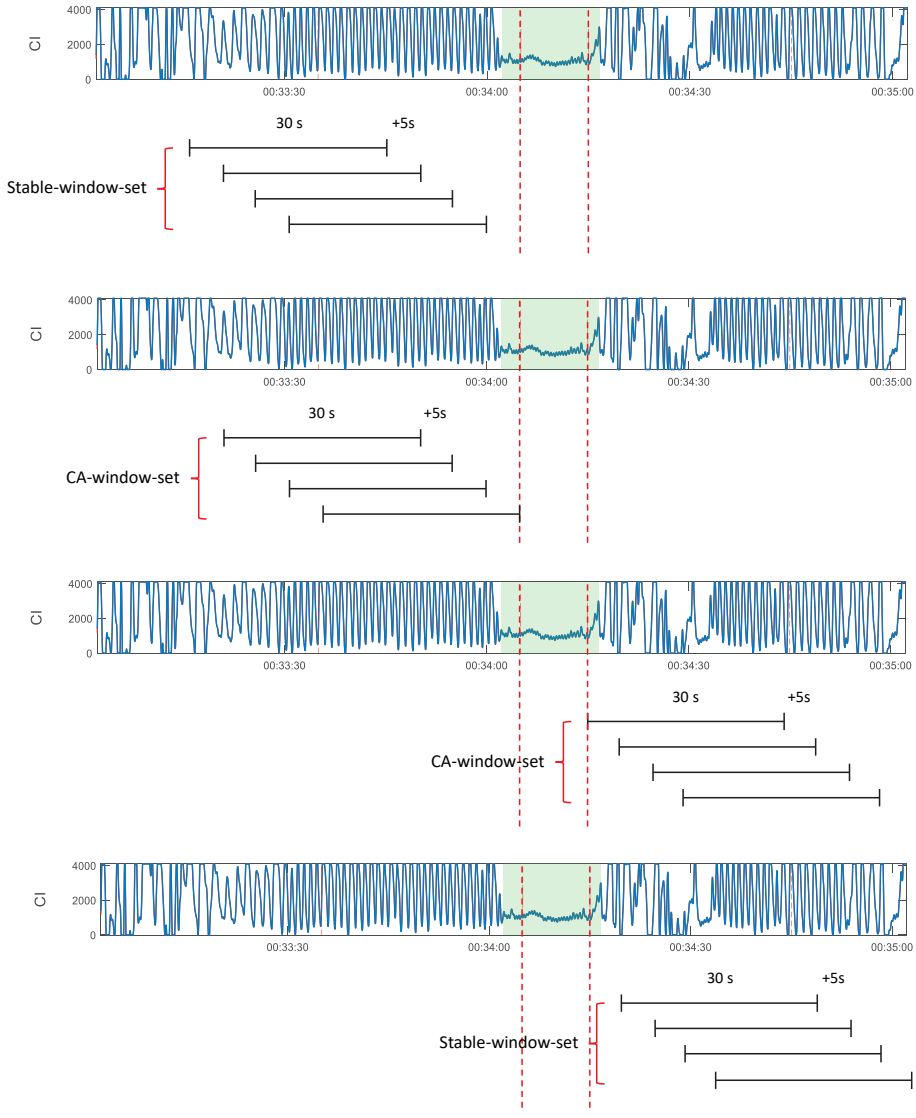


Figure 8.2: Annotation process for 4 consecutive moving windows (4-windows-sets) around a central apnea (CA) annotation.

b. a dataset created with a new KNN-based undersampling method, which only considered the distances within stable-windows-sets (i.e., majority class) and was therefore named Within-Majority undersampling. This was designed to keep a maximum of one stable-windows-set in groups of close windows-sets and to keep at the same time stable-windows-sets which were found very distant from all the others. This undersampling method was implemented for each patient

independently from the others and is shown in **Figure 8.3**. At first, it required the use of KNN to find for each stable-windows-set the 5 closest stable-windows-sets (i.e., 5 pairs of 'associated sets'). The matrix of distances for all pairs of associated sets in one patient was then sorted in ascending order. By proceeding from the associated sets with the minimum distance and by proceeding towards the associated sets with the maximum distance, for each pair of associated sets it was determined if one or both of the 4-windows-sets had already been seen before in previous associated sets. If so, both stable-windows-sets were discarded, if not, one was arbitrarily added to the dataset-to-keep. This process continued until the last 717 most distant pairs of associated sets (i.e., a minimum of 1 h of stable-windows-sets) were left and added to the dataset-to-keep. This whole process was repeated for the following patient.

In step 6, for both datasets resulting from step 5a and 5b we identified each cluster of consecutive stable-windows-sets. In case a cluster was found to be consisting of less than 4 consecutive stable-windows-sets, the stable-windows-sets immediately following the cluster were also included in the dataset. This step was performed in order to allow to perform a smoothing option over the annotations returned by the detection model based on machine learning, which is further described in **Paragraph 8.2.5**.

As a result of all the steps here described two datasets which included a total of 153564 4-windows-sets considering NearMiss 3 (approximately 213 h of data, consisting for the 6.69% of CA-windows-sets) and 203225 4-windows-sets considering Within-Majority (approximately 282 h of data, consisting for the 5.05% of CA-windows-sets) were defined to perform CA detection with machine learning.

8.2.4. Machine learning: detection model and cross-validation

An overview of the detection model based on machine learning and the cross-validation solution used in this study is shown in **Figure 8.4**.

Logistic regression (LR) with the elastic net penalty was selected since it is faster than many other algorithms (an important characteristic when large datasets are considered), it performed best when performing CA detection in our previous study [26] and it allows for an easier interpretation of the results for clinicians compared to other machine learning algorithms.

Leave-one-patient-out cross-validation (LOPO CV) was used since it allowed to always have all 4-windows-sets from one patient left as a test set, a solution that suits well with clinical practice applications, where each new patient admitted in a NICU is treated as a

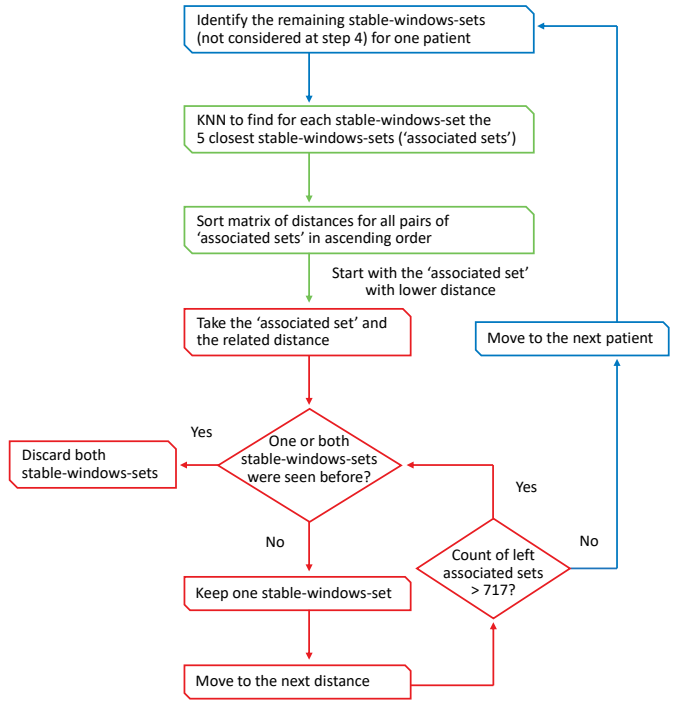


Figure 8.3: Overview of the Within-Majority undersampling method. At first, for each patient the stable-windows-sets are identified. KNN is then used to find for each stable-windows-set the 5 closest stable-windows-sets (i.e., 5 pairs of 'associated sets'). Next, the matrix of distances is sorted in ascending order. Starting from the associated sets with the minimum distance and continuing up to the associated sets with the maximum distance, for each pair of associated sets it was determined if one or both of the 4-windows-sets had already been seen before in previous associated sets. If so, both stable-windows-sets were discarded, if not, one was arbitrarily added to the dataset-to-keep. This process then process continued until the last 717 most distant pairs of associated sets were found and added to dataset-to-keep. The whole process was then repeated considering the following patient.

new test set. Differently from our previous study, hyperparameter optimization was performed by means of a grid search approach and a second CV separating the sets of windows from the remaining 9 patients into a training set (70% of the sets) and a validation set (30% of the sets). This allowed to extract the training and validation scores and to choose the optimal set of hyperparameters to be used on a test set based on two requirements: a set of hyperparameters leading to (1) a mean training score greater than a predefined threshold (Th_{class}), a solution implemented to prevent underfitting, and to (2) the minimum difference between the mean training and mean validation scores, a solution implemented to avoid overfitting. Th_{class} was initially empirically set equal to 0.9. In case less than 30% of the training scores computed for all sets of parameters were found to be greater than 0.9, its value was decreased each time by 0.01 until this requirement was met.

Two hyperparameters were optimized following this cross-validation: the parameter C, measuring of the strength of regularization (with lower C values leading towards stronger regularization), and L1 ratio, influencing the algorithm penalty oscillating from a mere L1 penalty to a mere L2 penalty (respectively with a value oscillating from 1 to 0).

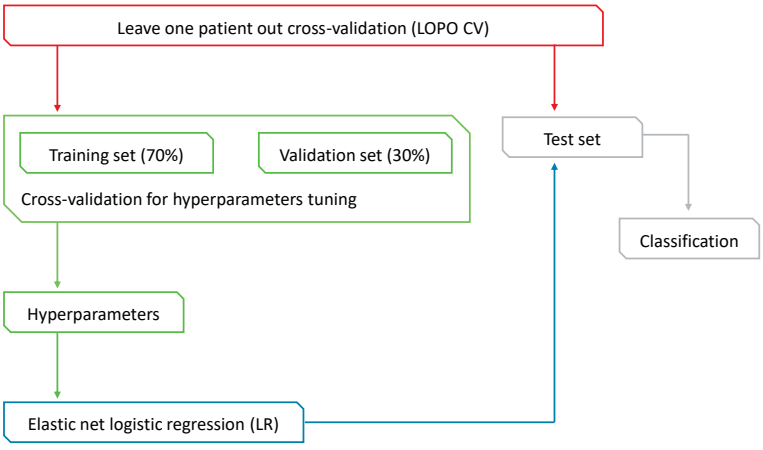


Figure 8.4: Overview of the detection model based on logistic regression (LR) with elastic net penalty and the cross-validation methods used in this study.

8.2.5. Machine learning: performance and effects evaluation

We applied the detection model described in the previous paragraph using the two datasets to extract the annotations for each 4-windows set (ML-annotations). In case these received a CA annotation, these were named CA-ML-annotations.

A first evaluation of the detection model used with the two datasets was performed using the area under the receiver operating characteristic curve (AUROC) as a metric. This allowed to identify the combination of detection model and dataset characterized by the highest mean AUROC to continue with all the following evaluations.

The second evaluation of the detection was performed by extracting the confusion matrices considering different thresholds for the false positive rate (FPR) in the mean ROC curve. In addition to this, the recall (TPR), precision (PPV), miss rate (or false negative rate, FNR) and the specificity (or true negative rate, TNR) were also extracted, in a similar way to what was done in our previous study [26]. An optimal threshold for a possible use in a clinical context was then selected and considered for the following evaluations.

The third evaluation was performed by computing the percentage of CA-ML-annotations per patient in each group of 4-windows-sets: pre-CAs, CAs, post-CAs, pre-Rejections,

Rejections, post-Rejections, real stable-windows-sets (i.e., stable-windows-sets that were added at step 5 of the pipeline described in **Figure 8.1**) and within-apneas (i.e., consisting of all stable-windows-sets included between consecutive CAs or Rejections which were separated by less than 2 min). We performed this evaluation 4 times: at first by considering single CA-ML-annotations returned by the detection model, and next by considering different smoothing options, which allowed to consider as true only CA-ML-annotations found for multiple (2, 3 and 4) consecutive 4-windows-sets. This evaluation allowed to have a better understanding of the overall distribution of false alarms in the whole dataset.

The fourth evaluation was performed by computing the percentages per patient of true CAs (and Rejections) annotated by the detection model. This was performed by considering how many CAs received a minimum of 1 CA-ML-annotation. At the same time, the percentages per patient of Rejections that received a minimum of 1 CA-ML-annotation were also computed to get a better understanding of the percentages of apneic events who were wrongly annotated as CAs (false alarms related to apneic events that were different from CAs). This evaluation was performed by considering the original ML-annotations as well as different smoothing options.

The fifth evaluation was performed by computing the percentages per patient of CA-ML-annotations around the start of the CAs and the Rejections annotated by the clinical experts. This allowed to verify when alarms were more likely to be triggered and was also performed considering different smoothing options.

8.3. Results

Results for first evaluation are shown in **Figure 8.5**, where the mean ROC curves computed for the two datasets are represented. A mean AUROC value equal to 0.81 and 0.85 were respectively found with the dataset that received NearMiss 3 undersampling and the one that received Within-Majority undersampling. Since the highest mean AUROC value was found with the dataset that received Within-Majority undersampling, results associated with it were considered in the following evaluations.

As a second evaluation different thresholds for the FPR in the mean ROC curve were used to extract the confusion matrices and the additional metrics. The results obtained with a FPR equal to 0.1, 0.2 and 0.3 were analyzed and the confusion matrices and the additional metrics obtained considering the selected FPRs as thresholds are presented in **Table 8.2**. A FPR equal to 0.1 returned a higher precision (i.e., fewer false alarms) compared to the other two thresholds but also led to a much lower recall (i.e., under

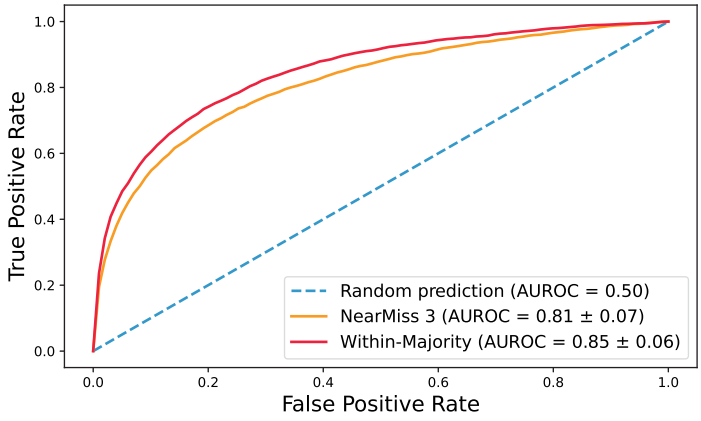


Figure 8.5: Mean ROC curves and AUROC values computed by using the detection model based on logistic regression (LR) with the elastic net penalty and the two datasets extracted considering NearMiss 3 and Within-Majority undersampling methods.

detection of CAs). A FPR equal to 0.2, characterized by a much higher recall, was considered as an optimal compromise between the two other thresholds and was therefore selected to perform the following evaluations.

As a third evaluation, the percentages per patient of CA-ML-annotations found for different groups of 4-windows-sets were computed and are shown in **Figure 8.6**. The percentage of correctly detected CA-windows-sets was found to be high (median percentage equal to 73.8%), however simultaneously high percentages of false alarms were generated (CA-ML-annotations found for different groups of 4-windows-sets which included pre-CAs, post-CAs, pre-Rejections, Rejections, post-Rejections, real stable-windows-sets and within-apneas). Smoothing options showed in **Figure 8.6B**-**D** progressively reduced the percentages of false alarms but also reduced the percentage of correctly detected CA-windows-sets.

As a fourth evaluation, the percentages of CA-ML-annotations per patient for CAs and Rejections, as annotated by the clinical experts, were computed and are shown in **Figure 8.7**. The red boxplots presented in this figure represent therefore the percentages of true CAs found per patient, whereas the orange boxplots represent false alarms related to apneic events that were different from CAs. Considering this evaluation, it can be noticed that even by considering a smoothing option performed with 3 or 4 consecutive ML-annotations the median percentage of CAs correctly annotated by the detection model would be high (equal to 88.3% and 84.0%, respectively) and the median percentage of Rejections wrongly annotated by the detection model would be

significantly reduced (reaching a median percentage of wrongly annotated Rejections equal to 48.4% and 42.6%, respectively).

As a fifth evaluation, the percentages per patient of CA-ML-annotations around the start of the CAs and the Rejections were computed and are respectively shown in **Figure 8.8** and **Figure 8.9**. It is clear that irrespectively of the choice of using the original CA-ML-annotations or any of the smoothing options, an increase in CA detection was found when more seconds of a CA were considered. In particular, using a smoothing option performed with 3 consecutive ML-annotations allowed to correctly detect a median of 74.2% of CA-windows-sets that overlapped with the first 25 s of a CA.

Table 8.2: Confusion matrices and metrics found considering a False Positive Rate (FPR) equal to (A) 0.1, (B) 0.2 and (C) 0.3 in the mean ROC curve computed using the detection model based on machine learning and the dataset obtained by means of Within-Majority undersampling.

A FPR = 0.1		Estimated Labels	
		Central Apnea	Reject
True Labels (Annotations)	Central Apnea	6602	3671
	Reject	19083	173869
Recall (TPR)	0.64	Precision (PPV)	0.26
Miss Rate (FNR)	0.36	Specificity (TNR)	0.90

B FPR = 0.2		Estimated Labels	
		Central Apnea	Reject
True Labels (Annotations)	Central Apnea	7930	2343
	Reject	38014	154938
Recall (TPR)	0.77	Precision (PPV)	0.17
Miss Rate (FNR)	0.23	Specificity (TNR)	0.80

C FPR = 0.3		Estimated Labels	
		Central Apnea	Reject
True Labels (Annotations)	Central Apnea	8721	1552
	Reject	57326	135626
Recall (TPR)	0.85	Precision (PPV)	0.13
Miss Rate (FNR)	0.15	Specificity (TNR)	0.70

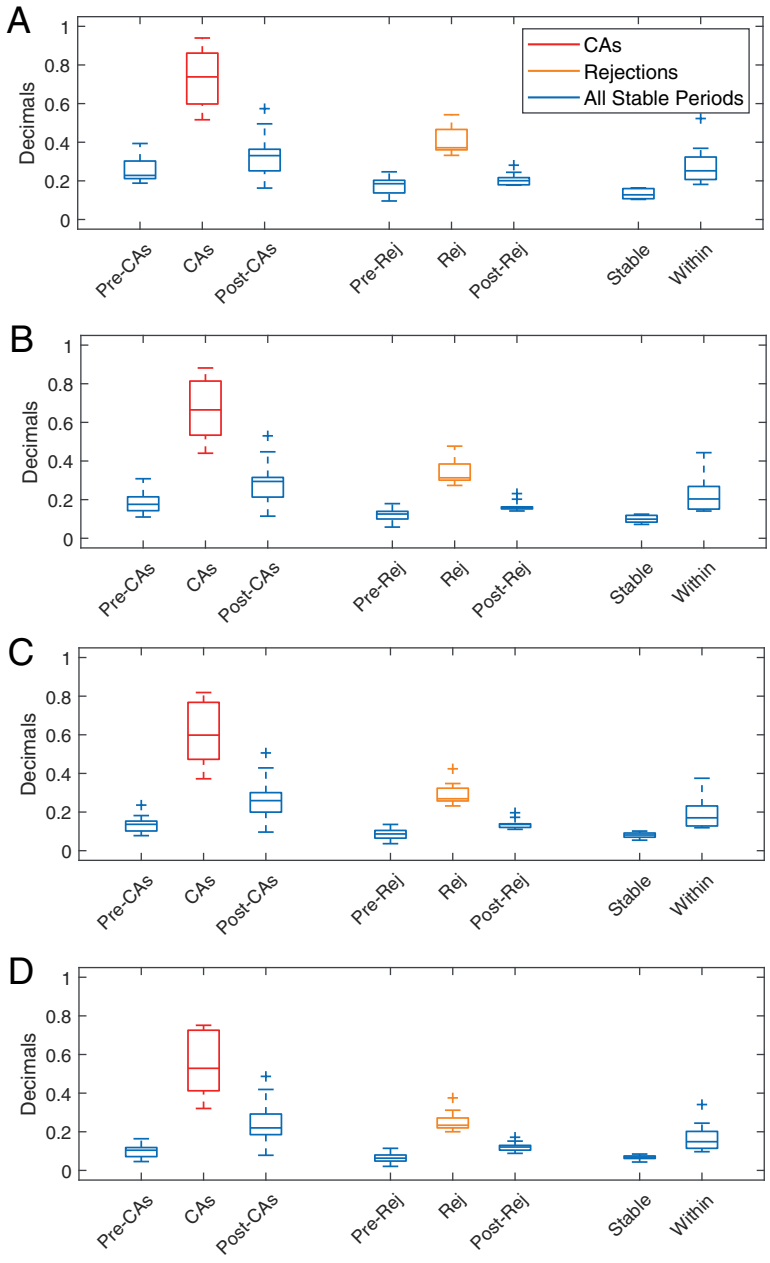


Figure 8.6: Boxplots for the percentages per patient of CA-ML-annotations found in different groups of 4-windows-sets: pre-CAs, CAs, post-CAs, pre-Rejections, Rejections, post-Rejections, real stable-windows-sets and within-apneas. Percentages were extracted by considering the dataset obtained with Within-Majority undersampling and by selecting FPR = 0.2 as threshold in the mean ROC curve. Percentages computed considering the original CA-ML-annotations (A) as well as a smoothing options with 2 (B), 3 (C) and 4 (D) consecutive CA-ML-annotations.

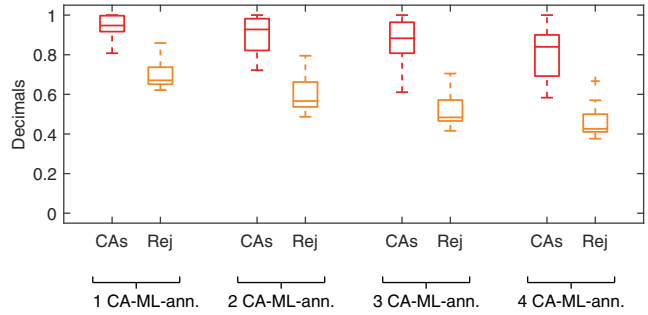


Figure 8.7: Boxplots for the percentages per patient of CAs and Rejections annotated by the clinical experts that received CA-ML-annotations from the detection model based on machine learning considering the dataset obtained with Within-Majority undersampling and a FPR = 0.2 as threshold in the mean ROC curve. The first boxplot from the left was computed by considering the percentages of CAs per patient that received a minimum of 1 CA-ML-annotation. The second boxplot from the left was computed by considering the percentages of Rejections per patient that received a minimum of 1 CA-ML-annotation. The following boxplots were instead computed by considering smoothing options with 2, 3 and 4 consecutive CA-ML-annotations.

8.4. Discussion

Apnea detection for premature infants, performed in the current clinical practice by means of standard monitoring techniques, is characterized by a high number of false alarms. Therefore, we decided to develop different strategies that can lead to better apnea detection and focused our attention on CAs. In our previous study we used a two-steps approach for CA detection that required the use of an optimized algorithm for the detection of apneic events from the chest impedance signal, followed by a CA detection model based on machine learning. In this study, we used machine learning to develop a one-step detection model to directly detect CAs in multichannel physiological data.

Since the full dataset was too imbalanced, two datasets were constructed out of all 4-windows-sets. A first undersampling method, NearMiss 3, which works by considering the distances between minority class samples and majority class samples, was specifically selected since a previous study showed that it can lead to a higher precision compared to other undersampling methods [38]. This was considered important since a reduction of false CA alarms would definitely provide benefits to NICUs in the form of less disturbances for the infants during sleep and a lower likelihood to experience alarm fatigue for nurses and clinicians. We also developed an additional undersampling method, Within-Majority undersampling, in order to consider the distances within majority class samples.

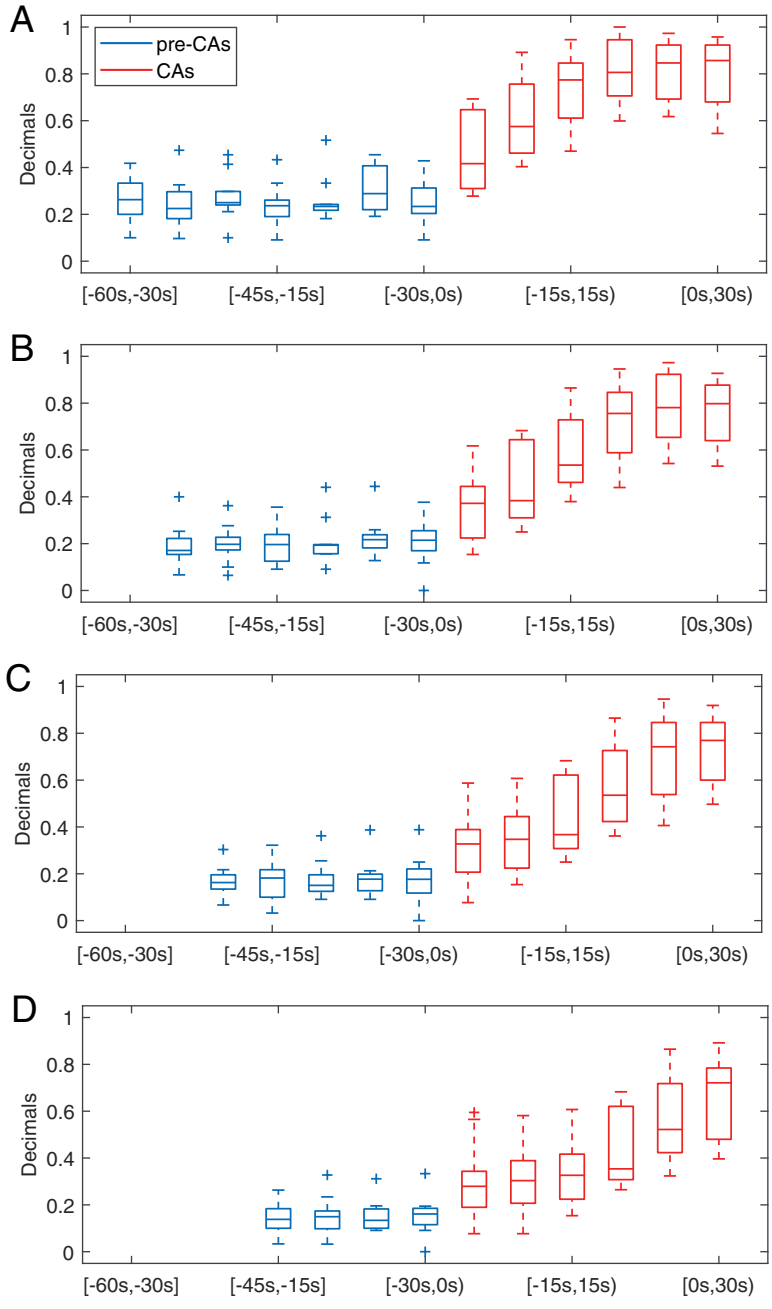


Figure 8.8: Percentages per patient of CA-ML-annotations found around the start of the CAs annotated by the clinical experts. These were computed by considering the original CA-ML-annotations (A) as well as a smoothing options with 2 (B), 3 (C) and 4 (D) consecutive CA-ML-annotations.

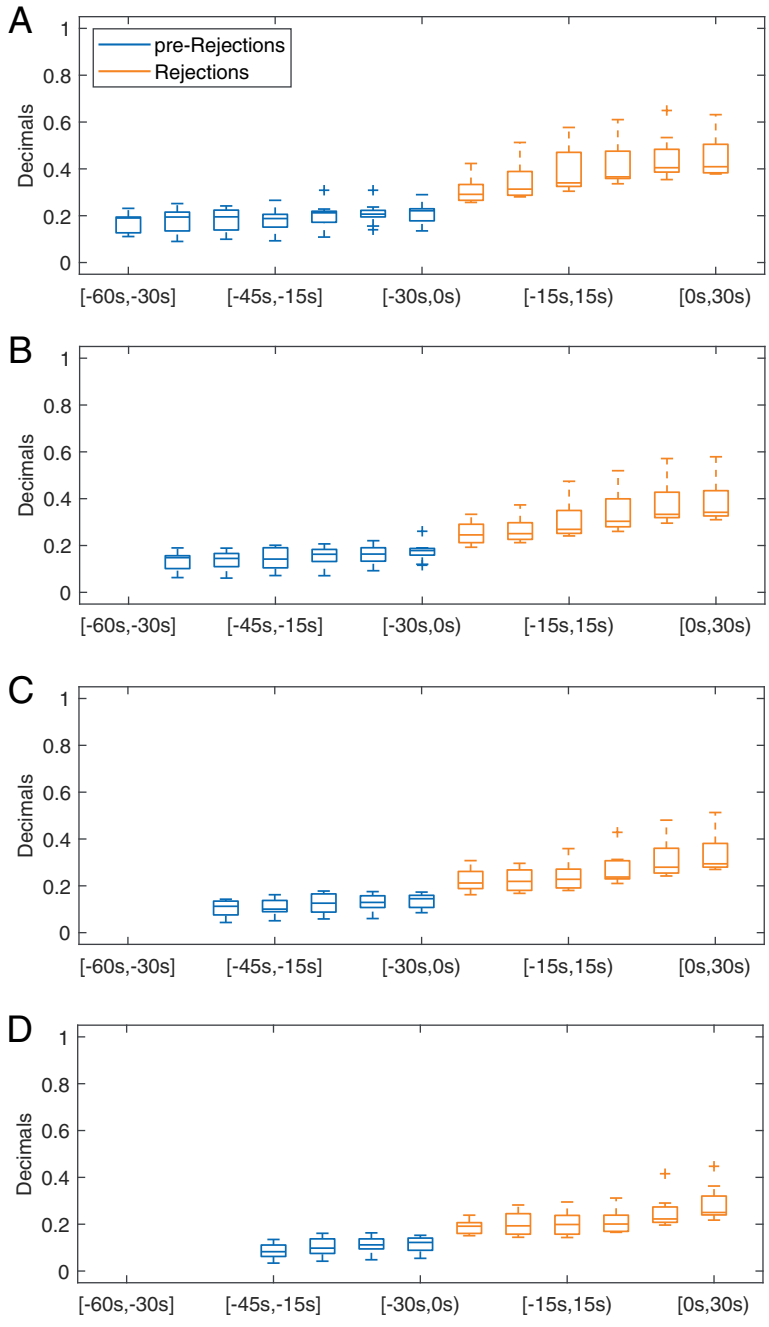


Figure 8.9: Percentages per patient of CA-ML-annotations found around the start of the Rejections annotated by the clinical experts. These were computed by considering the original CA-ML-annotations (A) as well as a smoothing options with 2 (B), 3 (C) and 4 (D) consecutive CA-ML-annotations.

In the first evaluation included in this study, our detection model based on LR with elastic net penalty showed mean AUROC values above 0.80 considering both datasets. These results were found to be in line with those obtained by the very promising study by Zuzarte et al., who performed the most extensive machine learning based apnea detection study to date by considering a population of 10 premature infants using continuous vital signs, and obtained a mean AUROC value equal to 0.77 [25].

AUROC values found in our study were also found to be in line or higher compared to those found in other studies that considered smaller datasets and used a wide variety of machine learning (e.g., decision trees, SVM, Bayesian adaptation of Gaussian mixture models) and deep learning models (e.g., Multi-Layer Perceptron, Stacked Auto-encoders) [21]–[24]. In our study, in particular, the highest AUROC value was found considering Within-Majority undersampling method and therefore the results obtained with this dataset were explored in further evaluations. We hypothesize that this undersampling method could allow for the inclusion a larger variety of samples since it keep majority class samples characterized by larger distances from the others, a selection that might therefore lead to better trained machine learning models.

In the second evaluation, confusion matrices and metrics were computed considering different FPR as thresholds. A FPR equal to 0.3 returned a precision equal to 0.13 and a recall equal to 0.85, metrics that are in line with those obtained by optimizing the Lee's algorithm for CA detection in a previous study performed by our group [20]. This result shows that it is feasible to perform CA detection by means of a multichannel detection model based on machine learning without necessarily considering the two-steps approach described in our previous study [26]. However, since in this study we wanted to increase the precision returned by the detection model, we chose to use a FPR equal to 0.2, which also led to a decrease in recall. Considering this choice, we then performed subsequent evaluations which were designed to better translate the results of the detection to the clinical practice.

In the third evaluation, we showed that a smoothing option, which allowed to consider as true only CA-ML-annotations found for 3 or 4 consecutive 4-windows-sets, allowed to reduce false alarms for stable-windows-sets. This idea of using these smoothing options was further strengthened by the findings in our fourth evaluation, where it was shown that a median percentage equal to 88.3% and 84.0% of CAs per patient received a CA-ML-annotation, respectively. At the same time the median percentage of Rejections per patient that received a CA-ML-annotation was reduced to 48.4% and 42.6%, respectively. We consider this an important improvement compared to the optimized algorithm for CA detection presented in our previous study, since this was characterized by a median precision equal to 19.7% [26].

Finally, in the fifth evaluation the percentages per patient of CA-ML-annotations around the start of the CAs and the Rejections showed that a smoothing option performed with progressively higher numbers of consecutive CA-ML-annotations led to a delayed detection of CAs. This delay is nonetheless preferable compared to the one introduced in the two-step approach by the optimized algorithm for the detection of apneic events since it is fixed and not influenced by the possible presence of upcoming apneic events, whose presence needs to be verified in order to perform the last step of the optimized Lee's algorithm for CA detection which is the main source of delay [26]. The delay found in this study is also considered to be reasonable considering that, according to most common definition of apnea of prematurity, several apneic events can only be annotated after 20 s of cessation of breathing [14], [17]. In addition to this, averaging times for alarms, working in a similar way to the smoothing options used in this study, are often used in the clinical practice in patient monitors [9] and our multichannel model is sufficiently fast to allow for these to be applied. Depending on the particular needs of each NICU, the smoothing option can be customized in order to prioritize either a further reduction in the number of false alarms (e.g., using a smoothing option performed using 3 consecutive CA-ML-annotations) or an earlier alarm onset. Irrespectively of the choice of using the original ML-annotations or any of the smoothing options, these results showed that detection of CAs is feasible with only few seconds of delay from their onset.

This study showed the feasibility of performing CA detection by means of the multichannel detection model based on machine learning. This allows for an easier implementation on bedside patient monitors since this one-step approach is faster, at least until no smoothing option is applied, compared to the optimized algorithm for the detection of apneic events, which requires several seconds of a cessation of breathing to return an apneic event to be then annotated by means a detection model. Our new multichannel detection model allows for a detection of CAs just a few seconds after their onset without additional time delays. In addition to this, different possibilities to tune the results are presented considering that different NICUs might prefer having a higher recall or a higher precision. The selection of FPR equal to 0.2 as threshold for the ROC curve and a smoothing option performed with 3 consecutive CA-ML-annotations, in particular, appear to be promising choices to significantly reduce the false alarms in stable periods (i.e., increase the precision) and at the same time to guarantee a high number of true alarms (i.e., minimize the decrease in recall).

The current study is however characterized by several limitations. The most important one is for sure the limited dataset that was used in this study, although it is still one of the largest datasets used for the development of an apnea detection model based on machine learning. This limitation is primarily due to the fact that annotating CAs in several hours of patient data is very time consuming. Future studies should consider

the inclusion of more patients and preferably validate the results on an external dataset from a different hospital.

Another limitation is that undersampling methods were applied also to the test sets and, due to time reasons, a detection on the complete test sets consisting of all 4-windows-sets for the included patients was not performed. The choice included in this study was however considered in order to compare similar groups of 4-windows sets used for both the training and the testing process in order to allow for a coherent transfer of information (i.e., to have a coherent distribution of 4-windows-sets in both the training and test dataset). Future studies will however need to verify the performance on more data, as well as including also the 10 patients characterized by a very low count of CAs that were excluded from this study.

Future studies should also investigate the effect of different undersampling methods, such as NearMiss 1 and 2, as well as oversampling methods, such as SMOTE, in order to verify their fairness [36], [38]. The Within-Majority undersampling, in particular, has been introduced in this study and further verification are needed. Nonetheless, a very similar performance in terms of mean AUROC value was found considering the dataset obtained with NearMiss 3 but the dataset that received Within-Majority undersampling also included several additional hours of patient data.

Future studies might consider the introduction of new features related to patient movements (e.g., features derived from wavelet-based algorithm), which performed very well to perform apnea detection as shown by Zuzarte et al. [25]. As shown in other studies focused on the prediction of other conditions occurring to premature infants, information related to the patient characteristics, such as GA, PMA and birth weight, could also play an important role in the detection in case additional patients will be included [39].

In this study, only a detection model based on LR with elastic net was considered due to computational reasons (i.e., LR is a much faster algorithm compared to others, such as SVM, when large datasets are considered), and since it performed very well in our previous study [26]. However, future studies would definitely benefit from the comparison with models based on different machine learning or deep learning algorithms.

One last limitation to be mentioned is that several false CA-ML-annotations are found before the start of a CA or a Rejection or immediately after their end. This is possibly the result of training with CA-windows-sets characterized by an overlap with a CA for only few seconds. Future studies might consider as CA-windows-sets only those 4-windows-

sets characterized by more seconds of overlap despite this solution might even cause a delay in the alarm onsets.

8.5. Conclusion

In this study we showed that CA detection in premature infants by means of machine learning is feasible in a time efficient way without preprocessing data, a characteristic which allows for an implementation on clinical patient monitors. To perform this study, we used two undersampling methods, including one that was developed to consider distances within majority class samples. This undersampling method performed well in generating a dataset that included a high variability in terms of included samples, which is typical for the clinical practice. CA detection using a multichannel detection model based on LR with elastic net returned a similar performance in terms of mean AUROC value compared to previous studies that investigated the use of machine learning to detect apneas. Further evaluations of the results obtained with the Within-Majority undersampling method allowed to identify a possible FPR threshold for the ROC curve and a smoothing option to be applied to the ML-annotations returned by the multichannel detection model to increase the precision without severely affecting the recall. These allowed to return a median percentage of CAs annotated per patient equal to 88.3% and a median of 74.2% CAs within the first 25 s from their onset.

References

[1] G. Schmalisch, "Neonatal monitoring," *Capnography, Second Ed.*, no. 2012, pp. 80–95, 2011.

[2] D. Ostojic *et al.*, "Reducing false alarm rates in neonatal intensive care: A new machine learning approach," *Adv. Exp. Med. Biol.*, vol. 1232, no. January, pp. 285–290, 2020.

[3] C. F. Poets, "Reducing alarms in the NICU," *Arch Dis Child Fetal Neonatal*, vol. 103, no. 4, pp. F297–F298, 2018.

[4] J. P. Keller, "Clinical alarm hazards: A top ten health technology safety concern," *J. Electrocardiol.*, vol. 45, no. 6, pp. 588–591, 2012.

[5] E. Institute, "Top 10 Health Technology Hazards for 2020 - Expert Insights from Health Devices," 2020.

[6] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomed Instrum Technol.*, pp. 46 (4), 268–277, 2012.

[7] C. Van Pul *et al.*, "Alarm management in a single-patient room intensive care units," *Recent Adv. Ambient Assist. Living – Bridg. Assist. Technol. e-Health Pers. Heal. Care*, pp. 119–132, 2015.

[8] C. Van Pul, H. Van De Mortel, J. Bogaart, T. Mohns, and P. Andriessen, "Safe patient monitoring is challenging but still feasible in a neonatal intensive care unit with single family rooms," *Acta Paediatr. Int. J. Paediatr.*, vol. 104, no. 6, pp. e247–e254, 2015.

[9] G. Varisco *et al.*, "Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU," *Acta Paediatr. Int. J. Paediatr.*, vol. 110, no. 4, pp. 1141–1150, 2021.

[10] C. McClure, S. Y. Jang, and K. Fairchild, "Alarms, oxygen saturations, and SpO2 averaging time in the NICU," *J. Neonatal. Perinatal. Med.*, vol. 9, no. 4, pp. 357–362, 2016.

[11] H. Lee *et al.*, "A new algorithm for detecting central apnea in neonates," *Physiol. Meas.*, vol. 33, no. 1, pp. 1–17, 2012.

[12] R. Joshi, D. Kommers, L. Oosterwijk, L. Feijs, C. Van Pul, and P. Andriessen, "Predicting Neonatal Sepsis Using Features of Heart Rate Variability, Respiratory Characteristics, and ECG-Derived Estimates of Infant Motion," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 3, pp. 681–692, 2020.

[13] A. H. Gee, R. Barbieri, D. Paydarfar, and P. Indic, "Predicting Bradycardia in Preterm Infants Using Point Process Analysis of Heart Rate," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2300–2308, 2017.

[14] E. C. Eichenwald and Committee on Fetus and Newborn, "Apnea of Prematurity," *Pediatrics*, vol. 137, no. 1, 2016.

[15] J. M. Di Fiore, R. J. Martin, and E. B. Gauda, "Apnea of prematurity - Perfect storm," *Respir. Physiol. Neurobiol.*, vol. 189, no. 2, pp. 213–222, 2013.

[16] J. M. Abu-Shaweesh and R. J. Martin, "Neonatal apnea: What's new?," *Pediatr. Pulmonol.*, vol. 43, no. 10, pp. 937–944, 2008.

[17] J. Zhao, F. Gonzalez, and D. Mu, "Apnea of prematurity: From cause to treatment," *Eur. J. Pediatr.*, vol. 170, no. 9, pp. 1097–1105, 2011.

[18] M. A. Mohr *et al.*, "Very long apnea events in preterm infants," *J. Appl. Physiol.*, vol. 118, no. 5, pp. 558–568, 2015.

[19] B. D. Vergales *et al.*, "Accurate automated apnea analysis in preterm infants," *Am. J. Perinatol.*, vol. 31, no. 2, pp. 157–162, 2014.

[20] G. Varisco *et al.*, "Optimized Detection of Central Apneas Preceding Late-Onset Sepsis in Premature Infants," in *Annu Int Conf IEEE Eng Med Biol Soc*, 2021, pp. 5463–5468.

[21] R. D. Shirwaikar, U. D. Acharya, K. Makkithaya, M. Surulivelrajan, and L. E. S. Lewis, "Machine learning techniques for neonatal apnea prediction," *J. Artif. Intell.*, vol. 9, no. 1–3, pp. 33–38, 2016.

[22] N. Mago, S. Srivastava, R. D. Shirwaikar, U. D. Acharya, L. E. S. Lewis, and M. Shivakumar, "Prediction of Apnea of Prematurity in neonates using Support Vector Machines and Random Forests," *Proc. 2016 2nd Int. Conf. Contemp. Comput. Informatics, IC3I 2016*, pp. 693–697, 2016.

[23] R. D. Shirwaikar, D. Acharya U, K. Makkithaya, S. M, S. Srivastava, and L. E. S. Lewis U, "Optimizing neural networks for medical data sets: A case study on neonatal apnea prediction," *Artif. Intell. Med.*, vol. 98, no. July, pp. 59–76, 2019.

[24] K. Lim, H. Jiang, A. P. Marshall, B. Salmon, T. J. Gale, and P. A. Dargaville, "Predicting Apnoeic Events in Preterm Infants," *Front. Pediatr.*, vol. 8, no. September, pp. 1–7, 2020.

[25] I. Zuzarte, D. Sternad, and D. Paydarfar, "Predicting apneic events in preterm infants using cardio-respiratory and movement features," *Comput. Methods Programs Biomed.*, vol. 209, p. 106321, 2021.

[26] G. Varisco *et al.*, "Central Apnea Detection in Premature Infants using Machine Learning," *Comput. Methods Programs Biomed.*, vol. 226, p. 107155, 2022.

[27] M. P. Griffin and J. R. Moorman, "Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis," *Pediatrics*, vol. 107, no. 1, pp. 97–104, 2001.

[28] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, "Heart rate variability Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, no. 17, pp. 354–381, 1996.

[29] L. Cabrera-Quiros *et al.*, "Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning," *Crit. Care Explor.*, vol. 3, no. 1, p. e0302, 2021.

[30] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, 2006.

[31] X. Long, P. Fonseca, R. M. Aarts, R. Haakma, and J. Foussier, "Modeling cardiorespiratory interaction during human sleep with complex networks," *Appl. Phys. Lett.*, vol. 105, no. 20, 2014.

[32] P. Indic, E. B. Salisbury, D. Paydarfar, E. N. Brown, and R. Barbieri, "Interaction between heart rate variability and respiration in preterm infants," *Comput. Cardiol.*, vol. 35, pp. 57–60, 2008.

[33] A. Bauer *et al.*, "Phase-rectified signal averaging detects quasi-periodicities in non-stationary data," *Phys. A Stat. Mech. its Appl.*, vol. 364, pp. 423–434, 2006.

[34] J. W. Kantelhardt *et al.*, "Phase-rectified signal averaging for the detection of quasi-periodicities and the prediction of cardiovascular risk and the prediction of cardiovascular risk," *CHAOS*, vol. 17, no. 015112, pp. 1–9, 2007.

[35] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy maturity in premature infants Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. Physiol. Hear. Circ. Physiol.*, vol. 278, pp. H2039–H2049, 2000.

[36] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1260–1284, 2019.

[37] M. Chiera *et al.*, "Heart Rate Variability in the Perinatal Period: A Critical and Conceptual Review," *Front. Neurosci.*, vol. 14, no. September, pp. 1–23, 2020.

[38] J. Zhang and I. Mani, "kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction," in *Proc. Int'l Conf. Machine Learning (ICML '2003), Workshop Learning from Imbalanced Data Sets*, 2003.

[39] Z. Peng *et al.*, "A Continuous Late-Onset Sepsis Prediction Algorithm for Preterm Infants Using Multi-Channel Physiological Signals From a Patient Monitor," *IEEE J. Biomed. Heal. Informatics*, vol. PP, pp. 1–12, 2022.

PART THREE

Societal impact, conclusion and future research

CHAPTER 9

Societal impact

9.1. Societal impact

Alarm fatigue in hospitals has been nominated one of the top 10 of health technology hazards by the ECRI Institute since 2012 [1]–[4]. In order to reduce risks associated with alarms, both technological and workflow improvements have to be established by caregivers, engineers and industry working together as a team [5]. The work included in this thesis illustrates innovations that have been developed to optimize neonatal patient monitoring and alarm management for clinical practice. This work is the result of the collaboration between Eindhoven University of Technology (TU/e), Philips Research and Máxima Medical Center (MMC). These institutions are grouped within the Eindhoven MedTech Innovation Center (e/MTIC) collaboration, whose objective is to create a fast track for research, development, and implementation of sustainable innovations that reach the clinical practice. Within e/MTIC, the ALARM project, which is funded by NWO-TTW (Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO, grant number 15345) focuses on the development of methods to optimize neonatal patient monitoring in NICUs. The multidisciplinary team that collaborated in this project allowed to achieve quantifiable results which can have a long-term impact not only on premature infants and their families, but also on nurses and clinicians as well as on industry. The potential impact of this work is discussed in the following sections. Knowledge utilization is also detailed.

9.1.1. Impact on premature infants and their families

All strategies applied in this thesis aimed at developing models and solutions to improve neonatal patient monitoring. This thesis focused on improving alarm management in NICUs (PART ONE) and on improving the detection of apnea of prematurity (PART TWO). A particular focus was dedicated to this condition since it occurs frequently, generates numerous alarms and can lead to severe consequences [6]–[8]. In addition to this, the detection algorithms for apnea of prematurity that are currently implemented in patient monitoring systems appear to be suboptimal for premature infants since they trigger several false alarms (up to 65%) [9], [10].

The reduction in alarms – which was achieved in PART ONE of this thesis – has a direct impact on the sound environment experienced by the patients and their families as well as on the alarm burden perceived by the nurses and clinicians working in the NICU [11]. Decreasing the number of alarms can be very easily achieved by setting broader thresholds for the optimal target range of different physiological signals. However, such a strategy would have detrimental effects on the premature infants, as critical conditions may be missed. Therefore, we developed an optimization method, by focusing on both technological means and workflow adjustments, that reduced alarms while keeping the patients in their optimal saturation targets. We guided the implementation by

monitoring it using a dashboard which will be addressed in more details in **Paragraph 9.2**. This method allowed to reduce the number of critical alarms by 30% while at the same time increased the time spent by the patient in the optimal saturation target by 35%. Other studies have shown that a more silent NICU reduces patient stress and promotes sleep and autonomic nervous system development [12], [13]. A lower number of alarms sounding in the patient room also reduces the stress experienced by the parents. New therapies and interventions, focusing on improving the patient and family experience during their NICU stay, were introduced in the NICU over the last years [13]. One promising therapy aims at improving the infants' environment by using music therapy interventions, performed by means of soft singing, humming, talking or rhythmic reading, to reduce stress and improve parent-child bonding [12]. However, the effects of these interventions are not easy to measure. Therefore, we implemented a quantification method using signal analysis based on the heart rate variability. Results indicated that the music therapy interventions led to beneficial effects for the autonomic regulation of the participating premature infants.

The impact of improved detection of central apnea (CA) (PART 2) and a consequent reduction in the number of the related false alarms can lead to an improved neurocognitive development even in long term, as it is known from literature [14]. However, while the detection algorithm and models used in this thesis retrospectively improved CA detection, these benefits still need to be externally validated before they can be implemented in clinical patient monitor systems by industry.

9.1.2. Impact on nurses and clinicians

The importance of improving the working environment has been underestimated for a long time and only recently, thanks to a higher awareness regarding the consequences of alarm fatigue, has received greater attention. With all the works included in this thesis we aimed at reducing the high alarm burden that characterizes the current clinical practice in order to improve the workflow in NICUs. This result was successfully achieved for instance in the study included in **Chapter 2**. An important indication that can be inferred from this work is that only a collaboration between nurses, clinicians and technicians working in a NICU and industry can aim at solving such a challenge. This was shared by industry with other users from other NICUs by means of a 'clinical evidence series' paper, which is also shown in **Figure 9.1**. In addition to this, this work also led to the development of a new tool for the visualization of the time spent by a patient (or multiple patients) in the different target ranges for a specific physiological signal, as discussed in **Paragraph 9.2**.

Whether externally validated, detection models based on machine learning shown in different chapters of this thesis also hold the potential to improve neonatal monitoring

since they could further reduce the number of false alarms in the clinical practice in the near future.

9.1.3. Impact on industry

The work included in this thesis resulted in three different outcomes for industry.

The first consists of graphical representations which were developed to support alarm management in the clinical practice. **Paragraph 10.2**, in particular, provides further indications regarding the methodology and graphical representations which were developed in a combined effort between university, hospital and industry. These were all invented in such a way that they can be easily implemented together, for instance in a dashboard displaying all relevant information in the monitors present at the nurses' central station or through an alarm management system. These methods and graphical representations have been transferred to Philips, industry partner of this project and a provider of patient monitors for NICU where the technology can be potentially implemented.

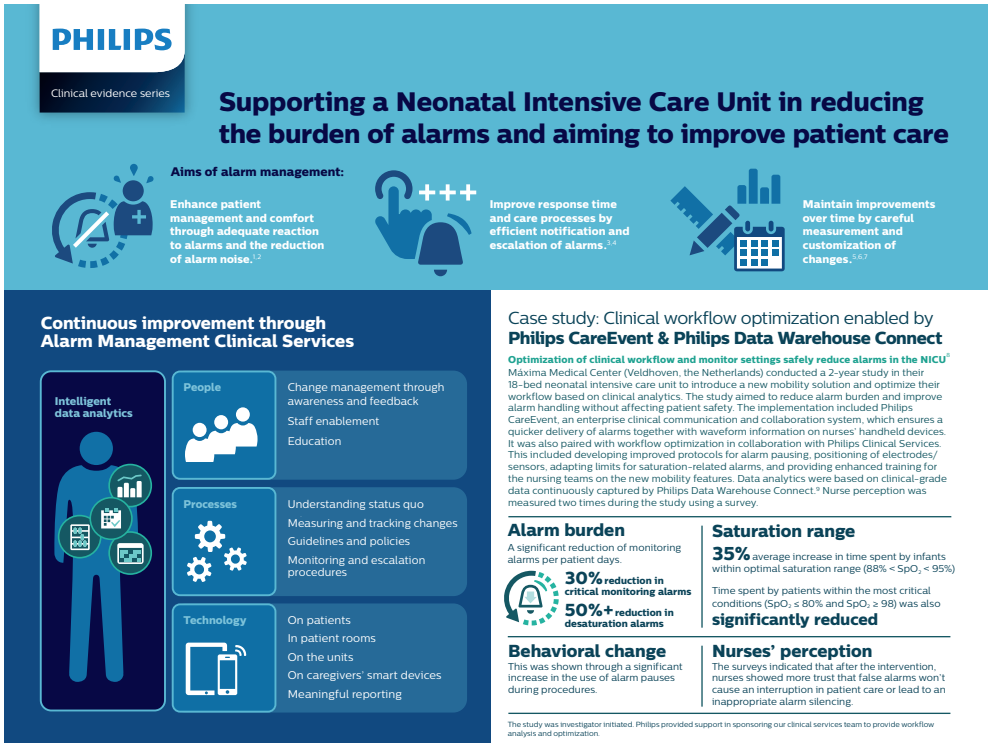


Figure 9.1: Clinical Evidence Series paper summarizing the results of the workflow optimization study included in Chapter 3. This paper is used by Philips to promote the alarm management solution named Philips CareEvent to other hospitals.

The second outcome is shown in **Chapter 2**, where it was shown that products and solutions can be combined with consultancy advice and have a larger impact on the clinical practice, as shown with a workflow optimization for the NICU of MMC. This work showed that the effects of the implementation of a proprietary product (namely Philips CareEvent, an alarm management solution) together with an optimization program showed the desired effect.

The third outcome includes the different optimized algorithms and models to perform detection of CAs included in this thesis. For their development, only information derived from patient monitors were used, therefore allowing for an implementation on current patient monitoring systems.

9.2. Knowledge transfer

Together with the works included in this thesis, three invention disclosures were generated. These are new visualization tools which were developed in close contact with Philips and the hospital personnel of MMC with the idea of finding innovative and simple solutions that could bring a benefit to both industry and hospitals. Together with the signal target area representation for the visualization of the time spent in the different target ranges for a specific physiological signal, which was presented in **Figure 2.3**, other visualization tools that were developed included: (1) a tool to visualize the alarm burden present in the different patient rooms of a hospital department considering its architecture, shown in **Figure 9.2**, (2) a tool to visualize alarm pressure in the NICU also considering alarm escalation, shown in **Figure 9.3**, and (3) a representation of the seconds of parallel alarms within a unit, an additional measure to investigate the alarm pressure, shown in **Figure 9.4**.

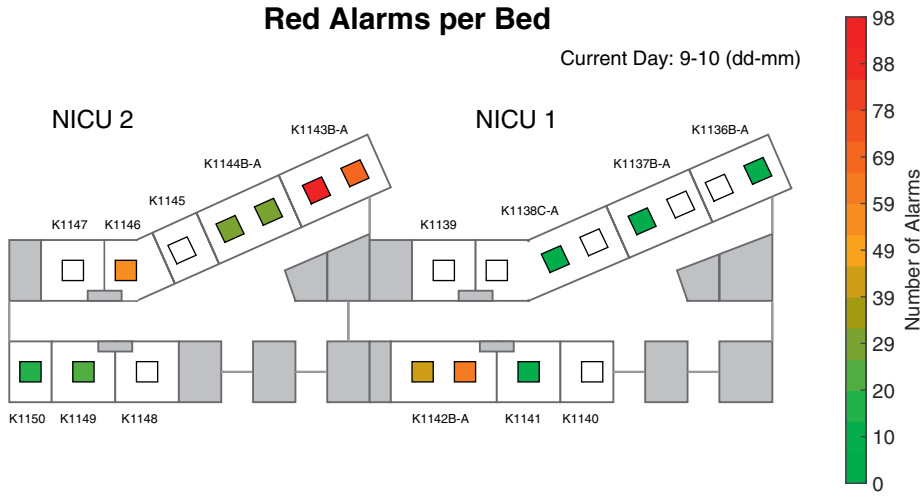


Figure 9.2: Tool to visualize the alarm burden present in the different patient rooms of a hospital department. For this representation, the architecture of MMC NICU is considered. Only red alarms per patient bed (indicated by colored squares in the different rooms) are included in the representation.

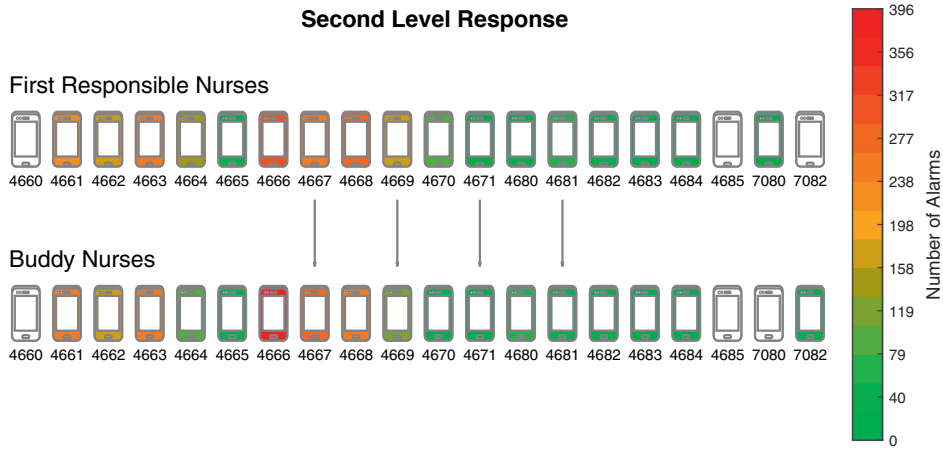


Figure 9.3: Visualization tool to provide an information about the number of alarms that are received by each nurse via their handheld device. In single-family room NICUs it is often possible that each nurse is associated to a buddy nurse, who is responsible to respond to the alarms escalated by the first (i.e., first responsible nurse in this visualization tool) in case a response to the patient's needs cannot be provided in time. In case both the first responsible nurse and the buddy nurse have to deal with patients characterized by important complications, this can cause further delays in their response and lead to alarm fatigue. This visualization tool was therefore designed for head nurses to monitor the level of pressure that each nurse is dealing with so that future nursing rounds could be better optimized.

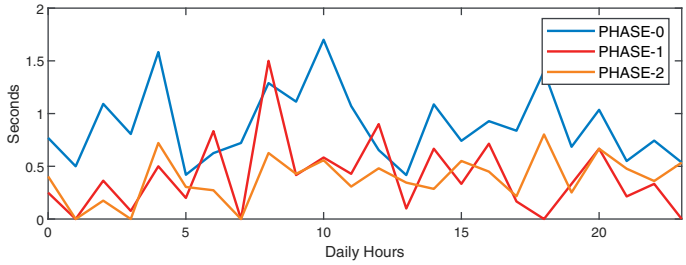


Figure 9.4: Representation of the seconds of parallel alarms measured in MMC NICU. Alarm occurrence in the form of seconds of parallel alarms was computed (for red alarms only) by considering for each hour all seconds in which a superposition between at least two alarms generated by different patients assigned to any nurse was found. All seconds of parallel alarms found within each hour were then summed and normalized considering the number of patients present in the unit. 24 values for each day (1 value per hour) included in the analysis were therefore extracted. The possibility to add an average representation for a time-period (as shown in this figure for three-months periods) was also added.

All these tools were developed considering the desire of the hospital personnel to introduce intuitive representations of both the patient conditions and the alarm burden in the NICU, which could be interpreted at a glance and based on the same color codes that characterize other medical devices already present in NICUs.

In addition to this, the work presented in **Chapter 2** also led to development of a dashboard for monitoring the process of alarm optimization, shown in **Figure 9.5**, which shows several information such as: the count of alarms, their duration and the time spent by patients in the optimal target range for oxygen saturation (SpO_2), measure which was included since deviations from the optimal saturation range could lead to severe problems for premature infants [14], [15]. However, the method is generalizable to other signals, like heart rate or blood pressure.

These tools were not filed as intellectual properties, since they are too general and not easy to patent. However, they were transferred to Philips for further use as they can be both implemented in the patient monitors or in the monitors located at the nurses' central station to provide additional information to the hospital personnel.

In addition to these, in PART TWO of the thesis a user interface for the annotation of CAs was developed in close contact with clinical experts to enable the time-consuming task of annotating CAs. This user interface provides visual information regarding all the physiological signals. Further details about this user interface are presented in **Chapter 5** and most precisely in **Figure 5.2**.

Finally, the calculated performances of the CA detection models based on machine learning models indicate where future resources may be invested in to reach the market with innovative solutions to improve the current clinical practice.

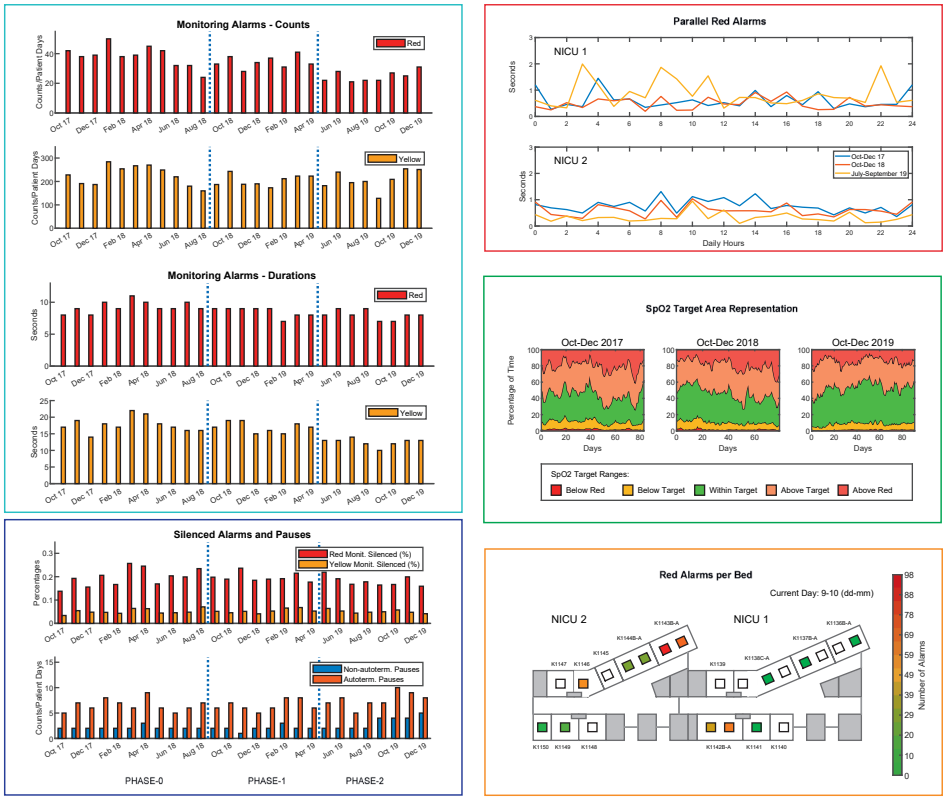


Figure 9.5: Dashboard for alarm optimization. This dashboard is customizable to include several different representations and measures. Here an example is provided, with 5 different information that are displayed. Part 1 (light blue box) of this example shows alarm distribution in the unit (count and duration of red and yellow alarms per patient per day). Part 2 (blue box) shows alarms that were silenced and paused (information available with percentages or counts). Part 3 (red box) show the seconds of parallel alarms calculated for different time periods. Part 4 (green box) shows the signal target area representation for the SpO_2 for different time periods. Part 5 (yellow box) shows the alarm burden present in the different patient rooms of a NICU.

References

[1] J. P. Keller, "Clinical alarm hazards: A top ten health technology safety concern," *J. Electrocardiol.*, vol. 45, no. 6, pp. 588–591, 2012.

[2] K. J. Ruskin and D. Hueske-Kraus, "Alarm fatigue: Impacts on patient safety," *Curr. Opin. Anaesthesiol.*, vol. 28, no. 6, pp. 685–690, 2015.

[3] E. Institute, "Top 10 Health Technology Hazards for 2016 - A Report from Health Devices," 2015.

[4] E. Institute, "Top 10 Health Technology Hazards for 2020 - Expert Insights from Health Devices," 2020.

[5] V. V. Paine, Christine Weirich Goel, E. Ely, S. Stave, Christopher D. Stemler, M. Zander, and C. P. Bonafide, "Systematic review of physiologic monitor alarm characteristics and pragmatic interventions to reduce alarm frequency," *J Hosp Med.*, vol. 11, no. 2, pp. 136–144, 2016.

[6] E. C. Eichenwald and Committee on Fetus and Newborn, "Apnea of Prematurity," *Pediatrics*, vol. 137, no. 1, 2016.

[7] R. E. Alvaro, "Control of breathing and apnea of prematurity," *Neoreviews*, vol. 19, no. 4, pp. e224–e234, 2018.

[8] K. Fairchild *et al.*, "Clinical associations of immature breathing in preterm infants: Part 1-central apnea," *Pediatr. Res.*, vol. 80, no. 1, pp. 21–27, 2016.

[9] O. P. Mathew, "Apnea of prematurity: Pathogenesis and management strategies," *J. Perinatol.*, vol. 31, no. 5, pp. 302–310, 2011.

[10] H. Lee *et al.*, "A new algorithm for detecting central apnea in neonates," *Physiol. Meas.*, vol. 33, no. 1, pp. 1–17, 2012.

[11] C. F. Poets, "Reducing alarms in the NICU," *Arch Dis Child Fetal Neonatal*, vol. 103, no. 4, pp. F297–F298, 2018.

[12] M. Filippa *et al.*, "Maternal Singing but Not Speech Enhances Vagal Activity in Preterm Infants during Hospitalization: Preliminary Results," *Children*, vol. 9, no. 2, pp. 1–10, 2022.

[13] D. R. Kommers *et al.*, "Features of Heart Rate Variability Capture Regulatory Changes During Kangaroo Care in Preterm Infants," *J. Pediatr.*, vol. 182, pp. 92-98.e1, 2017.

[14] A. Freudenthal, M. van Stuijvenberg, and J. B. van Goudoever, "A quiet NICU for improved infants' health, development and well-being: A systems approach to reducing noise and auditory alarms," *Cogn. Technol. Work*, vol. 15, no. 3, pp. 329–345, 2013.

CHAPTER 10

Conclusion and future research

10.1. Conclusion

Monitoring premature infants is extremely challenging [1], [2]. A huge variety of complications can occur to these patients every day due to their fragile conditions. Because of this reason, specialized care is provided to them in neonatal intensive care units (NICUs) [1]. The hospital personnel working in NICUs is composed of highly skilled nurses, clinicians and technicians who underwent years of training in order to provide timely assistance for all the possible needs that can arise [3]. The high standard of care found nowadays compared to the past years or decades were also supported by technological progress [4], which led to a continuous improvement of the medical devices applied for patient monitoring and treatment administration. On the other hand, the introduction in the patient room of several complex devices able to warn nurses and clinicians about possible complications led to a rise in the number of alarms sounding every day in NICUs [5]. Many of these alarms consist of non-actionable and false alarms [6], [7]. A high alarm burden not only affects patient sleep and development but can also result in alarm fatigue, with nurses and clinicians showing delayed and missed responses to critical conditions [8]–[11].

In order to make NICUs better environments for both patients, their families as well as nurses and clinicians, it has become a priority to develop new strategies to reduce non-actionable and false alarms without affecting the detection of critical conditions [12], [13]. At the same time, the development of innovative strategies to detect specific critical conditions, such as apnea of prematurity, may improve the monitoring of premature infants [14], [15].

The work included in this thesis aims at improving patient monitoring in NICUs. In particular, strategies to reduce alarms without compromising the patients' conditions are discussed in PART ONE of this thesis.

Chapter 2 shows that a reduction in the number of alarms found in NICUs can be achieved through an implementation of a new alarm management solution and a following optimization of the clinical workflow, revolving around a revision of the alarm settings and of protocols for sensor positioning, patient admission and discharge. This study, which was only possible thanks to a complete collaboration of the team of nurses, clinicians and technicians and important insights from industry, has in fact shown a significant decrease in the number of red (i.e., critical) alarms and in their duration. In addition to this, improved patient safety was found as shown by the significantly longer time spent in the optimal target range for the oxygen saturation (SpO_2).

Chapter 3 introduces the use of machine learning techniques to improve the NICU environment and proposes a new strategy to define whether an alarm is actionable. Different models to predict the need of nurses' response were developed by taking into consideration not only features extracted from the ECG and SpO_2 but also past interactions of nurses with the patient monitors (i.e., silencing of alarms and alarm pauses). This study showed promising results for instance by training only with features and nurses' responses extracted from stable periods and deteriorations and testing considering the complete dataset (results shows an AUROC value equal to 0.80 using the nearest mean classifier, resulting in a precision equal to 0.29, a specificity equal to 0.98 and a sensitivity equal to 0.34) but further studies need to prove this results in other contexts. Future studies focused on using nurses' responses to alarms could thus optimize the alarm chain.

Chapter 4 investigates whether new interventions can lead to reduced stress for the infants admitted to NICUs. The effect of music therapy interventions on the heart rate variability was measured by comparing features calculated from the patient monitor before, during and after the interventions. Results for different time-domain features (e.g., HR, SDNN and pNN50) showed lower variability during the interventions, indicating improved autonomic regulation. In addition to this, significantly lower values for the LF and HF Power were found during the interventions, which suggest improved autonomic regulation and less stress experienced by the infants.

PART TWO aims at providing new insights regarding central apneas (CAs) occurring during apnea of prematurity and at the development of new strategies for their detection.

Chapter 5 describes how a previously developed algorithm for CA detection was optimized considering a new dataset of 10 premature infants, by improving its precision without affecting its very high recall (very few apneas missed). The optimization of this algorithm was possible by designing a user interface which facilitated the annotation process for the clinical experts. A further application of the optimized algorithm on a second dataset consisting of 10 premature infants developing LOS followed by further annotations by clinical experts allowed to detect an increasing number of CAs closer to the onset of sepsis. This feature might be used in future works to predict the actual occurrence of this critical condition.

Chapter 6 aims at improving the understanding of mechanisms underlying apnea of prematurity by comparing CAs extracted from a population of premature infants with CAs simulated by means of a cardiorespiratory model. To perform this task, an existing model able to describe the cardiovascular system and its interaction with a baroreceptor

was improved to include the contribution of respiratory system and the gas exchange. This allowed to perform model simulations by varying different parameters such as the respiratory rate and alveolar compliance. Important similarities were found while comparing CAs from patient data and simulated CAs. These included a progressively deeper decrease in SpO_2 found for longer CAs and a recovery of the initial SpO_2 values which was found later in time. However, our model did not allow yet to simulate the observed decrease in HR following a cessation of breathing, which may be solved in a future model introducing chemoreceptors and lung stretch receptors.

Chapter 7 investigates the possibility of improving the detection of CAs by means of machine learning models, since apnea detection is currently affected by a high false alarm rate [16], [17]. All central apnea-suspected events (CASEs) (i.e., apneic events) were extracted from a population of 20 premature infants by means of the optimized algorithm described in **Chapter 6** and were further annotated by two clinicians to identify CAs. Different machine learning models were then trained using features extracted from the physiological signals (i.e., ECG, chest impedance and SpO_2) of the infants around the start of each CASE. All the models performed well in detecting CAs, as indicated by the high mean AUROC values (≥ 0.85). The best performing model, based on logistic regression with elastic net penalty and leave-one-patient-out cross-validation, proved to be particularly effective in correctly classifying CAs followed by bradycardia, which are considered the most critical type of CAs that can occur in premature infants. An important limitation that characterizes this study is that a two-step approach (i.e., a detection of CASEs followed by the use of a machine learning model for a detection of CAs) is required to perform CA detection. This approach is in fact slow since the optimized algorithm for CA detection returns a CASE only when several seconds of a cessation of breathing have already occurred.

Chapter 8 expands on the idea proposed in the previous chapter by using a multichannel model based on machine learning to perform CA detection also considering stable periods extracted from premature infants. This new model showed that it is possible to perform an accurate CA detection based on the signals directly, improving upon the low precision of the optimized algorithm for CA detection without significantly affecting the recall. In addition to this, this chapter showed the feasibility of turning the two-step approach of **Chapter 7** into a one-step approach that doesn't cause time delays and is faster for an implementation in clinical patient monitors for real-time use to detect CAs in physiological signals from premature infants.

10.2. Future research

The work included in this thesis presents new findings that require further investigation but also different limitations that can lead to further research. Some limitations are however particularly important since they did not affect the results obtained in this thesis but could prevent their reproducibility in future studies. Furthermore, other limitations were mentioned in multiple chapters within this thesis. Both these groups of limitations are therefore also listed in this paragraph together with possible solutions that can implemented in future studies.

First, the methods suggested in **Chapter 2** to perform an alarm optimization in MMC NICU and that revolved around the use of information derived from the alarms and the physiological signals (i.e., SpO_2) extracted from the patients might not necessarily lead to the same results in other NICUs. A decrease in the number and duration of the red alarms as well as an increase in the time spent within the optimal saturation range were in fact found in this study, but these results should be investigated in other contexts, where different NICU architectures, protocols and definitions for the alarms are present. In addition to this, the optimization that was considered for MMC NICU was performed for the red (i.e., critical) alarms only and not for the yellow (i.e., alerting) alarms. Future studies could therefore focus on the development of new strategies to reduce yellow alarms, which despite being characterized by a softer sound and a lower repetition rate can still affect the work of nurses and clinicians due to their frequent occurrence.

Second, in this thesis lots of attention was dedicated to the development of strategies to improve CA detection in order to reduce the high number of false alarms that occur in the current clinical practice. Future studies should however consider developing strategies to detect obstructive and mixed apneas as well, in order to provide a complete spectrum of solutions. In addition to this, while improved detection is one aspect, providing an early warning of critical situations and preventing the development of critical diseases accompanied by a higher CA rate (similarly to what was done within this group with the development of models aiming at an early detection of sepsis which allows for an early intervention [18], [19]) is also a relevant topic for further research.

Another important limitation is the size of the populations that were used in different chapters included in this thesis. This is important since observations or considerations found for a small group of patients may not be fully generalizable to a bigger patient group. An example for this is found in **Chapter 4**, where a total of 28 premature infants were analyzed for the effect of music therapy interventions on the heart rate variability. This limitation is however even more important when machine learning models are developed and therefore large datasets are needed. Relatively small datasets were used

in **Chapters 5** to **8** and consisted of populations of 10 up to 26 premature infants. This number was motivated by the high number of CASEs which needed to be annotated by clinicians to determine if these were CAs. In particular, for the population of 20 premature infants included in **Chapter 7**, a total of 7818 annotations were needed. The annotation process required therefore a considerable amount of time even if this was guided by a user interface which was specifically developed for this purpose. Including a bigger dataset in future studies, possibly consisting of several additional long CAs followed by desaturations, would allow to provide a better comparison with model simulations for this group of CAs in a similar way to what is presented in **Chapter 6**. In addition to this, since the ability of machine learning models to correctly detect, classify or predict events is very much dependent on the complexity that can be learnt from the training dataset, the use of bigger dataset could possibly lead to an increase of the performance of the models included in this thesis [20].

Another limitation which in machine learning studies is oftentimes paired with the limited sample size relates to the possibility of performing a validation of the results by means of a different dataset, possibly collected by another institution or hospital. This limitation is for instance present in **Chapter 3**, **7** and **8**, where only premature infants admitted to MMC NICU were included. Addressing this limitation for machine learning studies related to the medical field is not often possible mainly due to the privacy regulations regarding the access to patient data and the consequent limited number of public datasets that are currently available. In addition to this, the storage of this type of high temporal resolution data is not common as it requires special software and expensive storage. Furthermore, using raw sensor data might differ between manufactures and may require adjustments for the algorithms before a new dataset could be used. Standardization of the raw signals should be implemented and differences between manufactures should be investigated. Strengthening relationships between institutions in order to provide an easier access to shared datasets while at the same time complying with all the regulations set by the different countries could provide additional benefits to verify model performances on data that is acquired with different standards and monitoring devices from different vendors.

The use of only a limited set of machine learning algorithms for the development of the models included in this thesis is an additional limitation. The primary strategy which led to the selection of interpretable models to be used in these studies was the possibility to allow for an easier transfer of knowledge regarding the results, in order to reach a broader audience of readers. While this approach allowed to obtain very promising results related for instance to the idea of performing CA detection, as shown in **Chapters 5**, **8** and **9**, the use of deep learning models could possibly improve upon the results obtained until now.

In conclusion, the ambitious goal of improving alarm management in NICUs by reducing the alarm burden that characterizes the current clinical practice without compromising the patients' conditions was successfully achieved. The development of solutions to reduce the number of non-actionable and false alarms led to promising results but a further validation is needed before they can be implemented in clinical patient monitors or clinical alarm management systems.

References

[1] W. Chen, S. B. Oetomo, and L. Feijs, "Neonatal monitoring: Current practice and future trends.," in *Handbook of Research on Developments in E-Health and Telemedicine*, 2010, pp. 939–961.

[2] G. Schmalisch, "Neonatal monitoring," *Capnography, Second Ed.*, no. 2012, pp. 80–95, 2011.

[3] W. D. Barfield *et al.*, "Levels of neonatal care," *Pediatrics*, vol. 130, no. 3, pp. 587–597, 2012.

[4] J. Vincent, "Critical care – where have we been and where are we going?," *Crit. Care*, vol. 17, no. Suppl 1, pp. 1–6, 2013.

[5] S. Siebig, S. Kuhls, M. Imhoff, U. Gather, J. Schölmerich, and C. E. Wrede, "Intensive care unit alarms-How many do we need?," *Crit. Care Med.*, vol. 38, no. 2, pp. 451–456, 2010.

[6] R. Bonafide, Christopher P. Lin *et al.*, "Association between exposure to non-actionable physiologic monitor alarms and response time in a children's hospital," *J Hosp Med.*, vol. 10, no. 6, pp. 345–351, 2015.

[7] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomed Instrum Technol.*, pp. 46 (4), 268–277, 2012.

[8] B. D. Winters *et al.*, "Technological Distractions (Part 2): A Summary of Approaches to Manage Clinical Alarms with Intent to Reduce Alarm Fatigue," *Crit. Care Med.*, vol. 46, no. 1, pp. 130–137, 2018.

[9] R. M. Harris, J. Manavizadeh, D. J. McPherson, and L. Smith, "Do you hear bells? The increasing problem of alarm fatigue.," *Pa. Nurse*, vol. 66, no. 1, pp. 10–13, 2011.

[10] J. P. Keller, "Clinical alarm hazards: A top ten health technology safety concern," *J. Electrocardiol.*, vol. 45, no. 6, pp. 588–591, 2012.

[11] The Joint Commission, "Joint Commission Sentinel Event Alert: Medical Device Alarm Safety in Hospitals," *Sentin. Event Alert*, no. 50, 2013.

[12] S. Sendelbach and M. Funk, "Alarm fatigue: A patient safety concern," *AACN Adv. Crit. Care*, vol. 24, no. 4, pp. 378–386, 2013.

[13] K. R. Johnson, J. I. Hagadorn, and D. W. Sink, "Alarm Safety and Alarm Fatigue," *Clin. Perinatol.*, vol. 44, no. 3, pp. 713–728, 2017.

[14] R. Joshi, H. Van De Mortel, L. Feijs, P. Andriessen, and C. Van Pul, "The heuristics of nurse responsiveness to critical patient monitor and ventilator alarms in a private room neonatal intensive care unit," *PLoS One*, vol. 12, no. 10, pp. 1–14, 2017.

[15] D. Ostojic *et al.*, "Reducing false alarm rates in neonatal intensive care: A new machine learning approach," *Adv. Exp. Med. Biol.*, vol. 1232, no. January, pp. 285–290, 2020.

[16] S. B. Amin and B. Erica, "Monitoring Apnea of Prematurity: Validity of Nursing Documentation and Bedside Cardiorespiratory Monitor," *Am J Perinatol*, vol. 30, no. 8, pp. 643–648, 2013.

[17] H. Lee *et al.*, "A new algorithm for detecting central apnea in neonates," *Physiol. Meas.*, vol. 33, no. 1, pp. 1–17, 2012.

[18] Z. Peng *et al.*, "Body Motion Detection in Neonates Based on Motion Artifacts in Physiological Signals from a Clinical Patient Monitor," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 9, no. Mmc, pp. 416–419, 2021.

[19] Z. Peng *et al.*, "DeepLOS: Deep learning for late-onset sepsis prediction in preterm infants using heart rate variability," *Smart Heal.*, vol. 26, no. October, 2022.

[20] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Min.*, vol. 10, no. 1, pp. 1–17, 2017.

Acknowledgement

Thinking back about the recent years, I can recall several moments in which colleagues, friends and family members were available for me. Either for a simple talk to discuss how I was doing or for a technical discussion, several persons have been present, especially in moments of need. "Come, Mr. Frodo! […] I can't carry it for you, but I can carry you." said Samwise Gamgee to Frodo right on Mount Doom, when he was most in need of support. In a similar way to how Frodo thanked Sam at the end of their journey, I would like to thank all these persons for their help and support, which have been essential to reach the finish line of my PhD project.

First, I would like to start by thanking my main supervisor and promotor, Carola van Pul, who choose me for this PhD project in the first place. Without this choice as well as several indications given along the project, the results included in this thesis would have for sure not been possible. Her deep understanding of the field has in fact been crucial to guide me and to provide me with useful suggestions. I will also always remember the kindness that she showed in the moments in which I lacked trust in my work as well as her kind invitations to share some nice afternoons and evenings at her place.

I would also like to thank my two co-promotors, Peter Andriessen and Xi Long for their supervision along these four years. Their enthusiasm as well as the clinical and technical expertise have in fact inspired me and allowed me to develop a critical thinking about my work. Many thanks also to Zhuozhao Zhan for his availability and the precious technical advice that he shared along the project. Many thanks also go to prof. Ward Cottaar, which has been an important mentor in these years and always available for administrative problems. Thanks also to Rong-Hao for the interesting inputs during our technical meetings.

The colleagues and friends that I shared office with during these PhD years deserve a special thank. Thanks to Laura for the technical advice and for providing comparisons with her PhD trajectory, which allowed me to understand what I was doing good and where there was still room for improvement. Thanks to my good friend Zheng for the nice and casual talks, for the constant availability to discuss technical questions and for the support shown when I was most in need. Thanks also to Chunjiao and Jinjie for the time spent at the EMBC Conference in Glasgow and for the nice pool games and meals that we shared. Thanks to Ilde and Alessandro for the introduction to TU/e, for several lunches as well as for the nice moments spent together doing different activities such as playing futsal and having beers together. Thanks also to Alba, visiting PhD student from Spain, for the few but nice months spent working together in Eindhoven.

The hospital personnel from MMC that was involved in my project also deserves a special thank. Thanks therefore to Deedee for her cheerful and motivational attitude as well as for the help in the organization of my recent moving to my new apartment. Thanks to Heidi, for all the important insights that she shared regarding the organization of care in such an important NICU as the one in MMC. Thanks to the neonatologists Hendrik Niemarkt and Thilo Mohns, from whom I had the opportunity to learn a lot during different shared works. Thanks also to Jarno van den Bogaart, to Marieke Vervoorn and Krystal as well as all other nurses working in MMC NICU, who have always been very kind to me and available in case of need.

I would then proceed by thanking different researchers and employees working at Philips starting from Louis Atallah, Dirk Hueske-Kraus and Wilma van den Biggelaar, who provided relevant suggestions and expertise to perform the workflow optimization in MMC NICU and the related study. Thanks also to Sander Stuijk and Mohammed Meftah for the precious suggestions given during our regular meetings held at Philips Research during the first couple of years of my PhD.

Different students who were involved in different projects along these years also deserve to be mentioned. These include Ineke, Irene, Minette, Femke, Marieke and Lisan, who contributed to several nice studies and shared with the rest of the group some nice lunches and talks. A special thanks goes to Robbie and Boudewijn, who also provided a big help with my recent moving.

I would also really like to thank different persons that I have met during these years spent working at my PhD project and that provided a valuable help. Thanks therefore to Filippo, Serena, Mirka, Francesco, Priscilla, Matthieu, Monica, Mattia and Sara for the nice introduction that I received when I arrived in Eindhoven, as well as for several nice evenings spent together. Thanks to Giorgia, a friend that I had pleasure to share some time with both at my previous university as well as at TU/e. Thanks to Michiel, Dutch teacher and very much available to help. Thanks also to Yngwie, friend met at my previous place, for all the nice moments that we shared. Thanks to Kinga for the nice lunches shared at TU/e and for her help in finding a new accommodation.

Several friends from Italy also require a special thanks for what we have shared along these years. Thanks to my big group of friends from my previous university which includes Marta, Tommaso, Maria Grazia, Martino, Marco, Petra, Carlo, Caterina M., Caterina G., Andrea R., Giulia, Luigi, Giulia V., Rachele, Alessio, Chiara, Renato, Elena, Andrea M., Elia, Caterina S., Andrea C. e Elisabetta. Thank you for your calls, your thoughts and the nice holidays shared along these years. A special thank goes to Andrea A. and Giovanni, who persisted the most in calling me, to Benedetto, which has also been very kind to visit

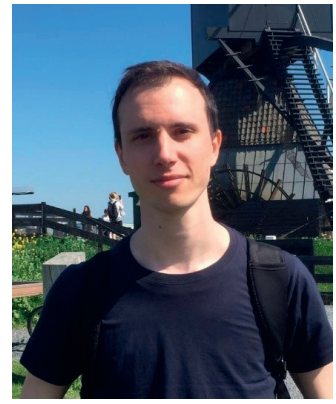
me often, and to Samuele, who drew the beautiful cover of this thesis. Thanks also to Gabriele, Fabrizio and Andrea P. for the nice talks and moments that we have shared in these years.

I would then proceed to thank my father, mother and sister, Ilaria, who have always been available in case of need and especially when I felt more in need. In addition to this, their precious suggestions have been crucial to face both some small and big problems found along the way. Thanks also to my cousins for the support shown in different moments.

At last, I want to thank a person who hold a special place in my heart that is my girlfriend, Ewa. She has in fact been a constant presence in my life in the recent years since the evening in which I met her almost by chance at a Greek restaurant in Eindhoven. Despite we are unfortunately far away from each other since almost two years, her caring attitude and interest towards me have been an important source of motivation to persevere and reach the end of this PhD project.

In conclusion of these long but dutiful greetings, I wanted to apologize to all those people whom I unfortunately forgot to thank: their advice and help were certainly no less important.

Curriculum vitae



Gabriele Varisco was born in Segrate (Milan, Italy) in 1990. Due to his interest in the scientific and medical fields, he enrolled at the university Politecnico di Milano (Milan, Italy), where in 2012 he was awarded the Bachelor's degree in Biomedical Engineering. During the Master's program he participated to the Erasmus program and followed different courses at Universiteit Gent (Gent, Belgium). Thanks to a joint collaboration between Politecnico di Milano and Columbia University Medical Center (New York City, USA), he then travelled to the United States to work at his Master's thesis, which was focused on the analysis of physiological signals extracted from newborn infants placed in different position and characterized by different sleep states. He was then awarder the Master's degree in Biomedical Engineering in 2015. After few years spent working as a clinical specialist in the surgery division of the company ab medica located in Cerro Maggiore (Milan, Italy), where he was responsible for the clinical training and technical support to surgeons in the entire Italian country for several surgical products, he moved to Eindhoven to start a PhD project in the Applied Physics department at Eindhoven University of Technology (TU/e, the Netherlands). His research is part of the Alarm-Limiting AlgoRithm-based Monitoring (ALARM) project funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek with grant number 15345. Máxima Medical Centre and Philips Research are the clinical and industrial partners of the project, respectively. His research interests include medical technologies, signal processing and the monitoring of physiological signals.

List of publications

Publications covered in this thesis:

- G. Varisco, H. van de Mortel, L. Cabrera-Quiros, L. Atallah, D. Hueske-Kraus, X. Long, E.J.E. Cottaar, Z. Zhan, P. Andriessen, C. van Pul, 'Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU', Acta Paediatrica, 2021
- L. Cabrera-Quiros, G. Varisco, Z. Zhan, X. Long, P. Andriessen, E.J.E. Cottaar, C. van Pul, 'Listen to the real experts: Detecting need of caregiver response in a NICU using multimodal monitoring signals', Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI'21), 2021
- G. Varisco, W. R. van der Wal, J. Bakker-Bos, D. Kommers, P. Andriessen, C. van Pul, 'Effect of Music Therapy Interventions on Heart Rate Variability in Premature Infants', 2022 44rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022
- G. Varisco, D. Kommers, X. Long, Z. Zhan, M. M. Nano, E.J.E. Cottaar, P. Andriessen, C. van Pul, 'Optimized detection of central apneas preceding late-onset sepsis in premature infants', 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021
- G. Varisco, I. Lensen, D. Kommers, P. Andriessen, P. Bovendeerd, C. van Pul, 'The effect of apnea length on vital parameters in apnea of prematurity - Hybrid observations from clinical data and simulation in a mathematical model', Early Human Development, 2022
- G. Varisco, Z. Peng, D. Kommers, Z. Zhan, E.J.E. Cottaar, P. Andriessen, X. Long, C. van Pul, 'Central Apnea Detection in Premature Infants using Machine Learning', Computer Methods and Programs in Biomedicine, 2022

Publications not covered in this thesis:

- G. Varisco, H. van de Mortel, L. Cabrera-Quiros, L. Atallah, D. Hueske-Kraus, X. Long, E.J.E. Cottaar, D. Kommers, T. Mohns, P. Andriessen, C. van Pul, 'Relevance of waveform information displayed on nurses' handheld devices in a single-family room neonatal intensive care unit', Pediatric Research, 2019
- L. Cabrera-Quiros, G. Varisco, Z. Zhan, X. Long, P. Andriessen, E.J.E. Cottaar, C. van Pul, 'Estimation of Heart Rate Directly from ECG Spectrogram in Neonate Intensive Care Units', 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society (EMBC), 2020
- Z. Peng, G. Varisco, X. Long, R Liang, D. Kommers, E.J.E. Cottaar, P. Andriessen, C. van Pul, 'A continuous late-onset sepsis prediction algorithm for preterm infants using multi-channel physiological signals from a patient monitor', IEEE Journal of Biomedical and Health Informatics, 2022

- Z. Peng, G. Varisco, R-H. Liang, D. Kommers, E.J.E. Cottaar, P. Andriessen, C. van Pul, X. Long, 'DeepLOS: Deep learning for late-onset sepsis prediction in preterm infants using heart rate variability', Smart Health, 2022