Bridging Human and Artificial Intelligence: Machine Learning, Data Platforms, and Decision Support Systems in Sleep Research

Benedikt Holm

March 8, 2025

Thesis Committee:

María Óskarsdóttir, Supervisor
Associate Professor, University of Southampton & Reykjavik University, United Kingdom & Iceland

Erna Sif Arnardóttir, Co-supervisor
Associate Professor, Reykjavik University, Iceland

Anna Sigríður Islind, Commitee member
Associate Professor, Reykjavik University, Iceland

Stefán Ólafsson, Commitee member
Assistant Professor, Reykjavik University, Iceland

Thomas Penzel, Commitee member
Professor, Charité University Hospital, Germany

Philip Terrill, Examiner
Associate Professor, University of Queensland, Australia

ISBN Print version: 978-9935-539-56-4
ISBN Electronic print version: 978-9935-539-57-1
Author's ORCID: 0000-0001-7213-2035

Contents

Contents **iii**

List of Figures **vii**

List of Tables **ix**

Disclosure on use of Generative Artificial Intelligence **xi**

Acknowledgments **xv**

1 Introduction **1**
1.1 Motivation 5
1.2 Outline and contributions 6

2 Related Work **11**
2.1 Adaptive segmentation 11
2.2 Unsupervised learning in sleep research 12
2.3 Digital platforms 13
2.4 Trust in AI 14
2.5 Decision support systems in sleep research 15
2.6 Research gaps 15

3 Methods **17**
3.1 Research approach 17
3.2 Data sources 18
3.3 Ethical considerations 19
3.4 Researcher's role 19

4 BreathFinder **21**
4.1 Introduction 21
4.2 Materials and methods 23

4.2.1 The BreathFinder algorithm . . . 25
4.2.1.1 Breath placement post-processing . . . 28
4.2.2 Data description . . . 31
4.3 Results . . . 32
4.3.1 Sensitivity analysis results . . . 33
4.4 Discussion . . . 34
4.4.1 Clinical implementation and applications . . . 37
4.4.2 Study limitations . . . 38
4.5 Conclusion . . . 38

5 VAE for sleep events **39**
5.1 Introduction . . . 39
5.2 Related work . . . 41
5.3 Methods . . . 42
5.3.1 Clustering analysis . . . 44
5.4 Experimental setup . . . 44
5.4.1 Dataset . . . 44
5.4.2 Preprocessing . . . 45
5.4.3 Data splitting . . . 46
5.5 Results . . . 47
5.5.1 Training . . . 47
5.5.2 Clustering analysis . . . 47
5.6 Discussion . . . 51
5.7 Conclusion . . . 53

6 Data ingestion framework **55**
6.1 Introduction . . . 55
6.1.1 Contributions . . . 58
6.2 Materials and methods . . . 59
6.2.1 Platform design . . . 60
6.2.2 Sleep technologist time and consensus validation . . . 63
6.2.3 Interviews with sleep technologists . . . 64
6.3 Results . . . 64
6.3.1 Platform performance . . . 65
6.3.2 Sleep technologist time and consensus validation . . . 65
6.3.3 Interviews with sleep technologists . . . 68
6.4 Discussion . . . 69
6.4.1 Main contributions . . . 69
6.4.2 Platform insights . . . 69
6.4.3 Clinical Acquiescence of AI . . . 70
6.4.4 Study limitations . . . 71
6.4.5 Future work . . . 71

6.5 Conclusion 72

7 ScoreCraft **77**
7.1 Introduction 77
7.2 Methodology 79
7.2.1 Platform 79
7.2.2 Recommendations 80
7.2.3 Data setup 81
7.2.4 Reference standard 82
7.2.5 Recruitment of participants 83
7.2.6 Study procedure 83
7.2.7 Analysis 84
7.3 Results 85
7.3.1 Participants 85
7.3.2 Aggregate analysis of scoring accuracy and time 86
7.3.3 Epoch-Level effects of recommendations on accuracy . . . 87
7.3.4 Epoch-Level effects of recommendations on decision-making time 91
7.4 Discussion 93
7.5 Conclusion 96

8 Discussion **99**
8.1 A unified view of AI in the field of sleep research 101
8.1.1 Designing AI with logical and physiological context 102
8.1.2 Standardizing and preparing data pipelines for AI 102
8.1.3 Enhancing clinical decision making with human-in-the-loop AI 102
8.1.4 Synthesis of the thesis 102
8.2 Towards theoretical applications in the field of integration and application of AI in the clinic 103
8.3 Towards practical applications in the field of integration and application of AI in the clinic 104
8.4 Beyond sleep research 105
8.5 Limitations 105

9 Conclusion **107**
9.1 Main Contributions 108
9.2 Future work 110

Bibliography **113**

A Appendices **131**

.1 ScoreCraft Study Completion Questionnaire 131

List of Figures

1.1 The three key areas of investigation that contribute to the central focus of AI in sleep research and medicine. 5

4.1 Phases of the respiratory cycle. 24
4.2 Respiratory Cycle Isolation algorithm flowchart. 25
4.3 Reference breath length histogram with model normal distribution. . 27
4.4 Key processes in the algorithm. (a) Detection merging procedure. (b) Overlap elimination process. 29
4.5 Visualisation of an example detection. 32
4.6 Sensitivity analysis of algorithm Recall and Precision to various parameters. (a) Window length, (b) Overlap percentage, (c) Correlation threshold, and (d) Probability threshold. 35

5.1 Variational autoencoder architecture components. (a) Encoder architecture. (b) Decoder architecture. 43
5.2 Example data of respiratory signals from BreathFinder. 46
5.3 Example of the reconstruction of the model. 47
5.4 Elbow analysis (distortion score is the sum of square errors). 48
5.5 Cluster prevalence for scorings. 48
5.6 Latent space sampling of individual breaths (a-z) with marked cluster centres (1-10). 49
5.7 Flow artefacts (green) vs. the population distribution (blue) in the latent space. 50
5.8 Hypopnea events (green) scored on the nasal airflow signal vs. the population distribution (blue) in the latent space. 50
5.9 Paradoxical breathing events (green) vs. the population distribution (blue) in the latent space. 51
5.10 Breaths drawn in the prone position (green) vs. the population distribution (blue) in the latent space. 51
5.11 Breaths drawn in the supine position (green) vs. the population distribution (blue) in the latent space. 52

6.1 Overview of the platform showing how the front end, processor, and splitter are combined. 61
6.3 Example of output 2 hours hypnograms for the n°1 PSG from 50 × 10 PSG, obtained using the processor and rendered in Nox Medical's Noxturnal software for manual review. 62
6.4 Front end user interface for uploading recordings. 66
6.4 Overlapped bars of scoring duration comparison of PSG with one sleep technologist using aSAGA-UA and the other two using the standard procedure. 73
6.5 Agreement analysis of 10 sleep technologists compared to a sleep technologist with or without aSAGA-UA assistance. The first column **(a,c,e)** shows the total agreement per polysomnographs. The second column **(b,d,f)** shows the agreement for gray area epochs tagged by the artificial intelligence. 74
6.6 A word cloud of the interview transcripts. 75

7.1 Map of sleep stages to deliberate misclassification for scoring recommendations. 80
7.2 Comparison of recommendations presented as human vs artificial intelligence (AI) 81
7.3 The MicroNyx scoring interface displaying a traditional PSG 82
7.4 Self-applied polysomnography in the MicroNyx scoring interface . . . 83
7.5 Sleep technologist change in accuracy between traditional and self-applied PSG. Each line represents one sleep technologist. 86
7.6 Sleep technologist change in decision-making time between traditional and self-applied PSG. Each line represents one sleep technologist. . . 87
7.7 Effect of recommendation presence on scoring session accuracy for traditional vs. self-applied PSG. 88
7.8 Effect of recommendation correctness on scoring accuracy. 89
7.9 Three-way line plot with grouped comparisons between effects of study type, recommendation presentation, and recommendation correctness on scoring accuracy. 90
7.10 Effect of recommendation presence on decision-making time for traditional vs. self-applied PSG. 91
7.11 Effect of recommendation correctness on decision-making time. . . . 92
7.12 Three-way Line plot with grouped comparisons between the effect of study type, recommendation presentation, and recommendation correctness on decision-making time. 94

8.1 The main focus areas of this thesis and their major contribution towards the central focus of AI in sleep research and medicine. 101

List of Tables

3.1 Data sources per publications. 18
3.2 Declaration of author contribution. 20

4.1 Evaluation results of the RCI algorithm. 32
4.2 Placement errors of the RCI algorithm. 33
4.3 Comparison between this work and related work. 36

5.1 Types of scoring events included in analysis. 45

6.1 Comparison of contributions of this work and similar work. 59
6.2 The layout of PSGs to be scored, where X indicates default automatic scoring and O indicates aSAGA-UA, that is aSAGA with gray areas. The numbers correspond to specific recordings in the 50×10 PSG. . . 63
6.3 Samples of processing time in minutes (min) taken by each queue according to file size in mego octets (Mo). 65
6.4 Fleiss's multi-rater κ mean±standard deviation estimated on overall hypnograms and gray areas epochs only by sleep technologists manually scoring and using aSAGA-UA assistance. 67

7.1 Generalized linear model linear regression results. The three-factor interaction term was included at first but was not significant. Thus, it was removed from the model. 90
7.2 ART ANOVA results for presentation, study type, and correctness on average decision-making time. 93

Disclosure on use of Generative Artificial Intelligence

During the writing of this thesis, a generative artificial intelligence (AI) (chatgpt.com) was used to proofread this document and as a recommender for improvements. ChatGPT was strictly used to improve text in terms of flow or grammar. The generative agent was never used for drafting or original writing for any chapter of this thesis, including the articles in chapters 4-7. AI-powered search engines (perplexity.ai, semantic scholar) were used in literature research. Finally, grammarly.com was used for grammatical correction, for which Reykjavik University provided an educational account.

Bridging Human and Artificial Intelligence: Machine Learning, Data Platforms, and Decision Support Systems in Sleep Research

Benedikt Holm

March 8, 2025

Abstract

Artificial Intelligence (AI) delivers groundbreaking automation capabilities to tasks that historically require manual human labor. However, its integration into fields like healthcare remains challenging due to concerns around interpretability, data standardization, and clinical trust. This thesis comprehensively explores AI's potential to enhance sleep medicine by addressing these challenges.

This work offers a holistic perspective on AI in sleep research, spanning the journey of data from collection and augmentation to its final presentation to human experts as well as the lifecycle of AI, from its inception to its integration into sleep medicine workflows.

The key findings include a novel respiratory cycle detection algorithm with 94% accuracy, insights into clustering respiratory events via unsupervised learning, and evidence that AI-assisted workflows reduce scoring time by up to 65 minutes while improving inter-rater agreement among sleep technologists. Furthermore, our research confirms that sleep technologists can work effectively alongside AI without significant distrust, highlighting a high level of clinical acquiescence.

The contributions focus on three key areas: (1) developing algorithms rooted in physiological principles to improve interpretability, (2) creating standardized data pipelines for scalable and reproducible AI deployment and (3) integrating human-in-the-loop solutions to enhance clinical decision-making.

These advancements underscore the transformative potential of AI in sleep medicine, providing a holistic view of its integration into clinical workflows. This research paves the way for the broader adoption of AI in healthcare by fostering trust, efficiency, and interpretability.

Keywords *Machine Learning, Sleep Research, Data Platform, Trust in AI, Decision support systems.*

Acknowledgments

I dedicate this work to the people in my life; without whom to share it, the load would have broken me a long time ago.

First, I would like to thank my fiancé, Matthildur Ármannsdóttir, who has been a consistent and endless font of support during the past eight years. My mother has supported my aspirations at every turn, and without her, I would not have even had the confidence to attempt a Ph.D, let alone finish it. My father, who has been an endless source of wisdom, calm, and motivation in the chaos and intense workloads. My younger brother, who has always been ready to lend a helping hand, no matter what.

I am lucky to be surrounded by the most incredible group of friends, who form such a diverse mosaic of personalities and backgrounds that I never found myself lacking perspective. There has never been a moment where I could not rely on their support and I count every time we meet as a blessing.

During my long stay at Reykjavík University, I have had the privilege of meeting and interacting with many wonderful individuals. Listing them all would make for an acknowledgments section longer than the thesis itself, but a few deserve special recognition: Sigríður Sigurðardóttir, Kenan Hoelke, Tiina Siilak, and Heiður Grétarsdóttir, whose expert knowledge and invaluable insights into the world of sleep have been instrumental in my journey.

The department of computer science at Reykjavik University is not only a fountain of peerless scholars but also outstanding mentors and role models, Marcel Kyas I want to thank personally for giving me a chance to gain my first teaching experience, and for convincing me to apply for the master's program. Anna Ingólfsdóttir has been a rock in the tumultuous and unpredictable journey and was one of the first, if not the first person to encourage me to apply for the Ph.D. program in the beginning.

I extend my heartfelt thanks to Jacky Mallett, with whom I could always confer, and for the privilege of teaching alongside her over multiple years.

Special thanks to the European Society of Sleep Technologists, under the leadership of Carlos Texeira, for providing invaluable assistance by allowing us to advertise the ScoreCraft study to its members.

I thank my thesis committee, Anna Sigríður Islind, Stefán Ólafsson, and Thomas Penzel, for elevating the thesis with their insightful comments and for the work we did together in the Sleep Revolution these past four years. Furthermore, I thank Philip Terrill for a thrilling thesis defense and for providing insightful analysis and conversation on the contents of the thesis.

Finally, I reserve my utmost gratitude for my supervisors, María Óskarsdóttir and Erna Sif Arnardóttir. Their patience, guidance, and resourcefulness were a guiding light in completing this work.

This thesis was funded by the European Union Horizon 2020 Research and Innovation Programme under Grant 965417. Funding to create a working prototype of the MicroNyx platform was received from the Icelandic Student Innovation Fund (Grant application #2311269-1101). Additionally, funds were received from the Erasmus+ study and/or traineeship mobility programme (KA131)

Gimli, Voodoo, Mosi, Pönnukaka

Chapter 1

Introduction

"It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to."

—J.R.R. Tolkien, The Fellowship of the Ring

Computer science has profoundly influenced nearly every aspect of human society, from facilitating near-instant communications via the Internet to significantly improving multiple industries with increased automation and data analytics [1]. By 2003, humanity had generated approximately five exabytes of data, a volume that, by 2013, was produced every two days [2]. This stark contrast between the immense benefits of computerization and the deluge of information generated by its ubiquity is almost paradoxical. The same machines that can simplify our lives and work also inundate us with data, some of which are not useful or informative.

Sleep medicine is a subfield within medicine that revolves around researching, diagnosing, and treating conditions and disorders that can influence sleep [3]. One such example is sleep-disordered breathing and obstructive sleep apnea for example, which are suggested to be considerably prevalent, having been found in 43.1% of the general population in Iceland [4], as well as estimated to affect one billion people worldwide [5], costing the United States $12.4 billion to diagnose and treat

in 2015 [6]. To get a diagnosis and then treatment for disorders like obstructive apnea, people have to undergo a sleep study, also called a polysomnogram (PSG), an overnight recording of various bio-mechanical signals such as electroencephalogram (EEG), airflow, respiratory inductance plethysmography, muscle activity, eye movement, cardiogram, and more [7]. A key role in the sleep medicine process is the sleep technologist or somnologist, whose expertise enables them to recognize various events during sleep, such as sleep stages, apnea, movements, tachycardia, or bradycardia, to name a few [8].

In contrast to sleep medicine, sleep research is the scientific study of sleep, in which researchers investigate sleep functions such as sleep stages and sleep-disordered breathing and employ academic methodologies to discover novel patterns in sleep and links between sleep events or sleep quality and day-to-day life. In many aspects, sleep research and medicine are conducted using historical approaches. For example, the scoring of sleep stages is done on a 30-second epoch basis because 30 seconds is the amount of EEG signal that researchers could fit on a sheet of paper [9].

After a PSG is collected, it must go through a manual scoring process, where an accredited sleep technologist goes through the entire PSG and labels it manually for sleep stages and other events such as obstructive apneas and arousal, to name a few [8]. This process has historically been time-consuming and can take sleep technologists up to a few hours to score a single recording. Additionally, there is considerable disagreement between sleep technologists; for example, when scoring sleep stages, there is only approximately 82.6% agreement [10], and even less for respiratory-related events when they occur, particularly for hypopnea (65.4% agreement) and central-apnea (52.4% agreement) [11]. Considering the laborious nature and the high disagreement rate between sleep technologists, it is evident that automating this workflow could drastically improve both efficiency and consistency.

Artificial intelligence (AI) refers to a subfield of computer science in which practitioners seek to use computational methods to solve practical problems that would otherwise require human intelligence to perform [12]. With the explosion of computing power brought on in the early twenty-first century [13], the field of machine learning (ML) has seen a surge in applications [14].

ML is a subfield of AI that revolves around creating programs to solve specific tasks without explicitly programming them, specifically by using statistical methods to optimize functions that map inputs to corresponding outputs [14]. ML is generally split into two methodologies depending on the solved tasks: supervised and unsupervised ML [15]. Supervised ML refers to models trained on input data to produce specific outputs. Supervised ML models create classifiers such as sleep staging or image classification models. This approach is ideal for cases where the data structure is well understood, and the association between the input and the output is intuitive and well-known [15]. The second category is unsupervised ML,

where the data output is unknown or otherwise not supplied to the model, causing the model to learn the data structures independently. Unsupervised ML is usually used for clustering analysis, anomaly detection, and dimensionality reduction [15]. For both categories of ML, the model slowly adjusts itself over many iterations to create a mapping between the input data and the output. This training can be time-consuming, but with the increased accessibility of high-powered computational tools, the speed at which models can be trained has risen drastically. After training, an ML model will always perform consistently, i.e., an ML model will always produce the same output for a given input. In contrast, a human may produce a different answer depending on the problem's complexity and factors such as exhaustion, inattentiveness, or inexperience [16].

AI has widely been used to perform tasks that traditionally require humans to perform [17], including in the healthcare industry [18], where AI is expected to enrich the sector [19] significantly. An essential issue with utilizing AI in healthcare is that the AI models can not be held accountable for an incorrect decision that negatively impacts a patient [20]. While significant work has been put towards making AI models solve problems in medicine, integrating AI into healthcare still faces various challenges [21]. Due to the requirement for accountability and human oversight, solutions that put the human expert at the forefront have been proposed for various applications, including the healthcare industry [22]. These human-in-the-loop approaches seek to leverage the automation capabilities of AI while still providing human expert oversight, preventing the AI model from producing incorrect or potentially harmful outputs [23].

Various tasks in the workflows of sleep technologists are time-consuming, with the manual scoring process being the most obvious example, often taking sleep technologists up to two hours to score a single PSG [24]. AI has been widely explored as a means of automating this process, with multiple automatic sleep staging [25] and apnea detection models proposed [26]. However, the adoption of AI in sleep medicine and research does not come without its challenges. Firstly, an alternate approach to the analysis of sleep stages has been proposed and is called 'adaptive segmentation' [27], which suggests placing the boundaries of sleep stages at points in the EEG, where the frequency components of the brain waves change, rather than on fixed, arbitrary boundaries. In respiratory analysis, this adaptive segmentation lends itself well to the approach of breath-to-breath analysis, as the respiratory cycle of inhalation and exhalation are the simplest phases of respiration. Applying AI models to individual respiratory cycles instead of the commonly used fixed-length segments has a twofold benefit; firstly, the AI model does not need to be able to handle respiratory signals out of phase, and neither does it need to learn to 'count' breaths.

Secondly, researchers and practitioners must carefully consider model design to successfully integrate AI into clinical workflows, as complex tasks can require large models which are expensive to train. A critical factor in this integration is

the trust that patients and healthcare professionals can place in AI [28]. Explainability plays a key role in fostering this trust, as black-box AI models often produce opaque decisions that are difficult for humans to interpret [29]. Adaptive segmentation methods have been proposed to enhance explainability and robustness in analyzing physiological signals, as they are modeled on the natural state changes in data, avoiding the pitfalls of more rigid fixed-timespan analytical methods [27], [30]–[32].

Currently, in research on automating tasks using AI, the focus is mainly on the model's accuracy when solving the task. An example of this in sleep research is the accuracy when scoring sleep stages. However, less focus has been directed toward the practical implications of employing these models. Questions such as how much time these models can save in the scoring process or how they might enhance the accuracy of sleep technologists in sleep staging when used as an assistive tool remain relatively unexplored.

Then, for AI to be effectively integrated into the workflow of sleep technologists, it is essential to evaluate the effects of its inclusion to ensure that the tools do not inadvertently cause adverse effects. Furthermore, the level of trust clinical professionals place in AI solutions and their outputs requires extensive analysis before AI can safely be integrated into the workflows.

Finally, much research has also been done into measuring the inter-scorer variability, i.e., the agreement between different sleep technologists when scoring the same recording. Existing research into exploring the source of the approximately 17% disagreement between sleep technologists is scarce, attributing it mainly to difficulty scoring particular epochs, availability of pre-scored events, or calculated variabilities [33]. Decision support systems (DSS) have been developed to deal with such uncertainties, yet many aspects remain unexplored in general and in the domain-specific context of sleep research.

This thesis employs the Action-Design-Research (ADR) methodology to systematically address the challenges mentioned above of integrating AI in sleep medicine [34]. ADR provides a structured approach for solving complex real-world problems via iterative design. ADR prioritizes collaboration with end-users and stakeholders and ensures that the developed solutions are technically robust, contextually relevant, and aligned with the needs of sleep technologists.

An ecosystem perspective is frequently employed to gain a holistic view of AI adoption, wherein sets of programs or services that act as interfaces between humans and AI are conceptualized as components of a broader framework involving diverse stakeholders such as healthcare practitioners, patients, and researchers [35], [36]. These technological ecosystems primarily consist of managed platforms that oversee and coordinate various aspects of the data lifecycle. Examples include platforms dedicated to data collection, platforms facilitating interactions between healthcare professionals and data, and platforms responsible for communicating results and diagnoses to patients. Although the study of computer ecosystems is

well-established, this thesis defines ecosystems as interconnected computational services or algorithms collaborating to deliver a more integrated and comprehensive service than could be achieved individually.

1.1 Motivation

Based on the current landscape in the field where AI and sleep medicine and AI intersect, we identified three overlaping main fields of study, which can be seen in Figure 1.1.

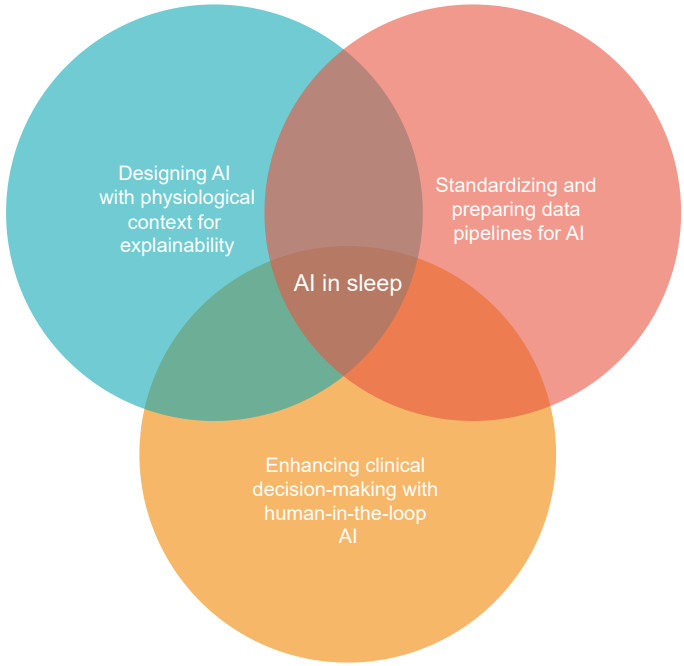


Figure 1.1: The three key areas of investigation that contribute to the central focus of AI in sleep research and medicine.

The first area of research, *"Designing AI with physiological context for explainability"* emphasizes the need for explainable AI in sleep research and medicine. By aligning AI decision-making with physiological and clinical logic, we can ensure that models are effective and interpretable for clinicians and researchers. This approach aims to build trust in AI by making its recommendations more transparent and grounded in established medical knowledge.

The second area, *"Standardizing and preparing data pipelines for AI,"* addresses the challenges of data collection, curation, and integration. Consistent and reliable data pipelines ensure that PSG data remains usable across diverse research settings and clinical applications. This area focuses on enabling scalability and reproducibility in automated workflows, which is critical for large-scale studies.

Finally, *"Enhancing clinical decision-making with human-in-the-loop AI"* explores the role of AI as a tool to support, rather than replace, sleep technologists. By incorporating DSS into scoring workflows, scoring accuracy can be improved, variability in scoring can be reduced, and workloads can be eased. This area also considers fostering trust and clinical acuity in such environments.

Together, these areas of investigation aim to address the broader challenges of integrating AI into sleep medicine and research, ensuring that technological advancements enhance research capabilities. From these areas of interest, we formulated three main research questions to which this thesis seeks to provide answers.

RQ1: *What could a fully-managed PSG ecosystem look like?*
RQ2: *How does one successfully integrate AI into sleep research or sleep medicine workflows?*
RQ3: *What are the effects of integrating AI into the scoring process?*

In the chapters ahead, this thesis seeks to provide convincing answers to these three questions and contribute meaningfully towards the three main areas of research outlined in Figure 1.1.

1.2 Outline and contributions

This thesis is aimed at practitioners and researchers working at the intersection of AI, decision support systems (DSS), and healthcare, focusing on sleep research and medicine. It provides valuable insights for academic researchers exploring innovative AI applications and healthcare professionals seeking to integrate AI into their workflows. Individuals in the sleep medicine and research community will find guidance on adopting AI-driven solutions to improve diagnostics and clinical workflows. For data scientists and ML engineers, this work highlights novel approaches to non-fixed-length data analysis, clustering methods, and pipeline optimization. This thesis bridges the gap between theory and practice by addressing critical challenges like trust, interpretability, and seamless AI integration, offering tools and strategies that advance these fields.

This section outlines the thesis structure and breaks down the following chapters, stating their subject and contributions towards the cumulative work this thesis represents.

Chapter 2

The related works section covers prior work in the fields this thesis seeks to contribute to and contextualizes the work in chapters 4 through 7. The chapter closes by outlining the main research gaps this thesis seeks to fill and why the gaps relate to the main questions introduced in chapter 1.

Chapter 3

This chapter details the overarching methodologies employed in writing this thesis, outlining the data used for each work in this thesis and an ethical statement. The chapter concludes by presenting an overview of the researcher's role, ending with the author's contribution statement.

Chapter 4

This chapter covers the design, development, and evaluation of BreathFinder, a novel breath detection algorithm. The purpose of the BreathFinder algorithm was to enable research into the respiratory system in a temporal context based on physiology.

The main contributions and findings of this chapter are as follows:

- We present an algorithm that detects individual respiratory cycles using a non-invasive, commonly collected signal during PSG.
- We thoroughly analyze the algorithm's performance using a hand-labeled data set containing multiple individuals and respiratory events.
- We found that the algorithm performs consistently at approximately 94% accuracy and recovers well from artifacts such as movements and respiratory events.

This work was published in

Holm B, Borsky M, Arnardottir ES, Serwatko M, Mallett J, Islind AS, Óskarsdóttir M. BreathFinder: A Method for Non-Invasive Isolation of Respiratory Cycles Utilizing the Thoracic Respiratory Inductance Plethysmography Signal. *Nat Sci Sleep.* 2024 Aug 21;16:1253-1266. doi: 10.2147/NSS.S468431. PMID: 39189036; PMCID: PMC11345460.

Chapter 5

This chapter details applying unsupervised ML to respiratory data and performing clustering analysis on various sleep events on a breath-by-breath basis. This ap-

proach utilized the aforementioned BreathFinder algorithm to separate the individual respiratory cycles and trained a variational autoencoder (VAE) to condense breaths into two variables.

The main contributions and findings of this chapter are as follows:

- We found that the VAE could encode and decode the breaths despite the considerable data compression.
- We found that some respiratory events tended to cluster together in the latent space over others.

This work was published in

B. Thordarson, A. S. Islind, E. Arnardottir, and M. Óskarsdóttir, "Exploration of Sleep Events in the Latent Space of Variational Autoencoders on a Breath-by-Breath Basis," *in Proceedings of the 56th Hawaii International Conference on System Sciences, Maui, Hawaii*, pp. 3091–30911.

Chapter 6

This chapter outlines the work towards designing a robust and scalable data ingestion platform that handles the receipt of PSG data and pre-processes it before finally augmenting it with automatic scoring. The effects of the augmentation on the scoring speed and accuracy were evaluated with the help of three sleep technologists. The main contributions and findings of this chapter are as follows:

- We found that AI assistance showed promise in lowering the scoring times, sometimes by as much as 65 minutes.
- We found that the AI assistance positively affected the scoring agreement, particularly for the less experienced sleep technologist.
- We introduce a new generalized term, 'Clinical Acquiescence,' which refers to human experts' willingness to accept AI's workflow assistance.

This chapter was published in

Holm B, Jouan G, Hardarson E, Sigurðardottir S, Hoelke K, Murphy C, Arnardóttir ES, Óskarsdóttir M and Islind AS (2024) An optimized framework for processing multicentric polysomnographic data incorporating expert human oversight. *Front. Neuroinform.* 18:1379932. doi: 10.3389/fninf.2024.1379932

Chapter 7

This chapter covers novel research where an online scoring interface was used to collect consensus scoring from 16 sleep technologists across Europe. Each sleep

technologist scored four hours of sleep for traditional and self-applied PSG, featuring recommendations varied in correctness and whether they were presented as being from a human sleep technologist or an AI.

The main contributions and findings of this chapter are as follows:

- We found that sleep technologists did not display a detectable difference in accuracy or scoring time between traditional or self-applied PSG.
- We found that the recommendation correctness affected the scoring accuracy significantly, with correct recommendations improving scoring accuracy. In contrast, incorrect recommendations hurt the accuracy of sleep technologists.
- Our results indicate that the sleep technologists displayed no biases in recommendation presentation, indicating a high level of clinical acquiescence for AI in sleep staging.

This work is in the submission process at the Journal of Sleep Research.

Chapter 8

The body of work is summarized in this chapter and the significance of the findings in chapters 4 to 7 is contextualized with the existing work and how the contributions tie together to provide a holistic methodology for AI in sleep research and medicine.

Chapter 9

This chapter finalizes the thesis, providing a closing statement and presenting the key contributions. The thesis concludes by providing suggestions for future work and how subsequent scientific ventures may build on it.

Chapter 2

Related Work

This chapter discusses existing literature and describes the gap in research this work intends to fill. The chapter is divided into four main sections, one for each article in chapters 4 through 7.

2.1 Adaptive segmentation

A limited amount of literature exists on using adaptive segmentation to identify individual breaths or respiratory cycle isolation (RCI), with existing methods mainly based on statistical analysis of signals such as peak and valley detection in the airflow, thoracic or abdominal RIP signals, feature extraction and modeling which are mostly derived from the audio signal recorded during the study [37]–[45].

Rosenwein et al. introduced a breath detection algorithm based on a random forest approach [40]. They derived 351 features from audio recordings and trained the model to detect inspirations and exhalations, reporting an 87% and 76% accuracy in predicting inspiration and expiration events, respectively. Yaha & Faezipour trained a support vector machine to detect respiratory phases using audio from a microphone placed in front of the participant's nose [41]. They report a 95% accuracy but acknowledge that the results depend on several factors of the microphone's placement and quality.Palaniappan et al. developed a neural network solution for classifying respiratory phases using respiratory sounds [46]. While they reported excellent overall performance for their model, they did not disclose its accuracy for the different breath phases. Hsiao et al. designed an attention-based autoencoder to perform respiratory segmentation on the audio signal [45]. They report a 91% accuracy in detecting the respiratory phases. A similar method applying adaptive segmentation using a variable window size was also proposed by Lalouani et al. with a novel system [47], which is capable of segmenting the audio signal into individual breath phases as well as classifying chronic obtrusive

pulmonary disease with a reported 92% accuracy in the detection of breaths vs. non-breaths with their model. Hult et al. proposed a bioacoustic method that can accurately time respiration from tracheal sounds using a summation method over the frequency domain of the audio signal. They evaluate their algorithm on 2074 respiratory cycles from two groups of participants, one being recorded in a quiet environment and another with acoustic disturbances from surrounding activities. They report detecting respiratory phases with 99% accuracy on participants in the group with the quieter environment and approximately 90% accuracy in the group recorded with more noisy conditions [44]. Alshaer et al. present a method for segmenting respiratory audio from a cardioid microphone but do not explicitly state performance metrics [48].

Moyles & Erlandson proposed a non-parametric statistical approach to RCI, based on detecting changes in the trend of the airflow signal, but do not provide any validation of their algorithm [37]. Korten & Haddad presented a pattern recognition algorithm that detects respiratory events in a barometric pressure signal [49]. They claim that the difference in mean values for inspiratory time, expiratory time, and total respiratory cycle time between the manually calculated and automatically detected values using the pattern recognition algorithm is minimal (<6%) but do not explicitly state performance regarding detections.

Lopez-Meyer et al. presented an RCI algorithm based on peak and valley detection on RIP signals to determine the beginning and end of a breath segment, reporting 96% precision when detecting breath cycles for participants during rest[39]. A Python library, RespInPeace, for RIP belt analysis is also available, which uses a peak and valley location algorithm to find respiratory cycles during a conversation but appears to have no published validation [42].

2.2 Unsupervised learning in sleep research

Korkalainen et al. [50] designed and trained a combination of a convolutional neural network and a long short-term memory (LSTM) neural network to classify sleep stages using a single frontal EEG channel. They found that classification accuracy decreased with increased obstructive sleep apnea syndrome severity. Cen, Yu, Kluge, *et al.* [51] achieved 79.6% accuracy when classifying one-second signal segments into normal, obstructive apnea, and hypopnea classes. They used deep convolutional neural networks to automatically learn the features of the airflow, oxygen saturation, and RIP signals. Thommandram, Eklund, and McGregor [52] showed that the k-nearest neighbor classifier could achieve up to 91.2% accuracy in classifying one-minute signal segments as apneic or non-apneic when using modest features derived from the respiratory inductance signal. This paper used statistics of breaths marked by peak detection but did not derive breath-specific features. Rosenwein, Dafna, Tarasiuk, *et al.* [53] classified respiratory events as

apneic/hypopnea versus non-apneic/hypopnea on a breath-by-breath basis using a random tree method. They used intuitive respiratory features such as breathing rate and ”duration to last respiratory event” for the classification and reported achieving 86.3% accuracy when classifying breaths as apneic/hypopnea. Nikkonen et al. [54] achieved an 88.9% accuracy on classifying OSA events using an LSTM network, with an average error in the apnea-hypopnea index of 3.0.

Outside medical approaches, VAE has been used for tasks such as anomaly detection [55], text classification [56], and recommender systems [57]. In the medical field, VAE and other AE methods have been applied in various ways, mainly for the electrocardiogram (ECG) [58]–[60], but also with some applications for the electromyogram (EMG) and EEG [61]. Costa, Sánchez, and Couso [58] applied a novel class-dependent implementation of the VAE for detecting atrial fibrillation, improving the latent space clustering by simultaneously training a classifier on the latent space with the VAE. Pastor-Serrano, Lathouwers, and Perkó [62] proposed an adversarial VAE trained to generate and classify respiratory signals during training using the latent space to classify baseline shift breathing irregularities. They trained the adversarial VAE on fixed-length signal segments collected by a stereotactic radiosurgery device.

2.3 Digital platforms

Digital platforms, as an area of research within information systems, are software solutions that facilitate the connection to key infrastructures of institutions via controlled collections of software or services, allowing for value-creating interactions between internal resources and external consumers or producers [63]. Many of the industry’s leading giants, such as Google, Apple, Facebook, Amazon, and Microsoft, have pioneered the design of digital platforms in healthcare [64], with digital platforms regarded as being drivers of technological adoption in the healthcare industry [65]. Digital platforms are increasingly utilized for a greater variety of tasks, such as collecting research data in the form of data collection platforms [66] and assisting human experts in various tasks such as DSS [67]. In sleep, digital platforms have been utilized to integrate heterogeneous sleep data through a dynamic digital platform and data pipeline [68]. They are well posed to facilitate communication between healthcare professionals and patients [69].

Platforms can act as enablers of DSS by acting as the intermediaries between the decision-making mechanisms and the humans interacting with the DSS [70]. Providing sleep technologists with automatic assistance during the scoring process has significantly reduced PSG scoring time, with some work showing improvements by factors of 1.26 to 2.41 [71]. Moreover, automatic sleep scoring models have been observed to halve the scoring time [72], [73].

Rayan, Szabo, and Genzel presented challenges and advancements in auto-

matic sleep scoring in the context of rodent and human sleep research [24]. They noted limitations in handling atypical data and a lack of flexibility but also noted that automation can make the process more efficient. A recent study found that deep-learning-based automatic scoring software strongly correlated with manual scoring in sleep staging and the apnea-hypopnea index while significantly reducing scoring time, thereby improving workflow efficiency in sleep laboratories [73]. Oxholm et al. interviewed nine healthcare professionals and five patients about their attitudes towards using data from electronic health records in an algorithm to screen for alcohol abuse in hospitals [74]. Professionals were mixed in their views, appreciating the tool's time-saving potential but concerned about losing instinctual decision-making. While this work is only tangentially related to ours, the authors highlight the requirement to include healthcare professionals in integrating automatic algorithms. Gerla, Kremen, Macas, Dudysova, and Mladek presented a computer-assisted approach for sleep staging using EEG recordings and AASM 2012 scoring rules, focusing on actual clinical data with artifacts and missing electrodes, evaluating the influence of AI in clinical settings by comparing traditional manual sleep stage classification with AI-based methods, including expert-in-the-loop strategies, for the analysis of EEG recordings in sleep studies [75]. In a later study, Gerla et al. [76] developed a semi-supervised method for evaluating PSG, blending expert-scored segments with automated classification. This approach, tested on healthy individuals and chronic insomnia patients, showed enhanced efficiency and accuracy in sleep data analysis compared to conventional manual scoring methods, demonstrating the impactful role of AI in streamlining sleep study workflows.

2.4 Trust in AI

For AI to effectively integrate into professional workflows, a certain level of trust in its outputs is essential. Without this trust, end-users may misuse the technology or refuse to engage with the AI tools in their work [77]–[79]. Many large AI models operate as so-called "black box machines" with opaque and not easily understood decision-making processes, hindering users from trusting their outputs [80]. Moreover, measuring trust in AI presents a challenge, as it is often subjective and not always aligned with the actual reliability of the tool being evaluated [79]. Conversely, humans can also exhibit overreliance on AI, sometimes making counterintuitive decisions based solely on AI suggestions, even when these contradict their own assessments and available information [81]. Such overreliance can potentially lead to catastrophic outcomes [82]. Detecting overreliance is inherently difficult [83], but it can be mitigated through strategies like providing explanations for AI outputs [84] or encouraging cognitive engagement with the AI system [85].

A significant factor that affects the level of trust in AI that healthcare profes-

sionals express is the transparency or explainability of AI, with models becoming more 'opaque' in their decision-making as they gain complexity, leading to an increased academic interest in explainable AI (XAI) [29]. XAI has been the focus of various research over the years, and the field of XAI study suffers from multiple challenges, such as a need for a unified concept and lack of standard terminology, as well as the tradeoff explainability tends to have on model accuracy [86]. Despite these challenges, XAI has succeeded in improving clinicians' ability to interpret the decisions made by AI models in various tasks, along with increasing trust in the machine learning model [23].

2.5 Decision support systems in sleep research

As stated previously, automatic suggestions are able to potentially save a significant amount of time in the scoring process, [71]–[73]. Some work exists on implementing DSS for sleep that use ML methodologies, but ultimately fail to take the last step and measure the impact of the automatic scoring models on the workflow of sleep technologists [87]–[90]. Liang et al. designed a clinical DSS for sleep staging, where a rule-based decision tree first scored the recordings. A reliability voting system flagged low-reliability epochs, which were then re-scored by sleep technologists, saving significant time by only focusing sleep technologists' attention on low-reliability epochs [72]. Similarly, Bechny et al. used the U-Sleep algorithm [91] to score sleep stages and trained a confidence network to find uncertain epochs. They report that to achieve a Cohen's Kappa of above 90%, less than 29% of epochs needed to be reviewed, significantly lowering the workload of sleep technologists. Hwang et al. designed and evaluated a clinical DSS that improved the macro-F1 score of less-experienced sleep technologists from 56.75 to 60.59 by presenting automatically detected features such as sleep spindles, but was noted by senior technologists to lack sufficient specificity in its presented information [92].

2.6 Research gaps

The relevant research gaps we identify and seek to address in this thesis are varied: While RCI algorithms have been proposed and described, there is a distinct lack of rigor in evaluating the algorithms, leading to either absent, unintuitive, or lackluster accuracy measurements. Additionally, the impact of common events such as movements or respiratory events such as apnea on the proposed algorithms is not commonly included in the evaluation when it is present. Another problem with some approaches is that formal algorithm validation is often not provided on patient data. Issues with equipment, wide variations in patient behavior, and noisy environments such as partner breathing or background noise may not have been adequately explored.

The currently most-explored application of AI in predicting sleep stages or respiratory events in the analysis of arbitrary fixed-length epochs such as 30 seconds, one minute, and five minutes, to name a few. Such fixed-length epoch approaches present an obvious problem, as events, particularly respiratory events, are variable in length, and any physiological or biomedical process is unlikely to be fully represented by or reliably captured in arbitrary fixed-length so-called epochs.

While studies have been performed exploring the practical factors of integrating AI into the workflows of sleep technologists, studies that explore the trust and clinical acceptability of AI are few. Many studies also base their research on aggregate-level analysis, with few exploring more granular, epoch-based behaviors. To our knowledge, studies exploring potential prejudice or bias against AI as a scoring assistant have also not been produced.

Chapter 3

Methods

3.1 Research approach

Throughout the work of this thesis, the methodology most closely followed was the action-design-research (ADR) methodology [34]. ADR addresses two major challenges: 1) identifying and addressing a problem situation encountered in a real-life situation and 2) creating an artifact to solve the problem situation identified and evaluating said artifact. ADR was deemed a good fit for this thesis due to its iterative and emergent nature, which posed it as a good fit for addressing the interdisciplinary challenges at the intersection of AI, healthcare, and sleep research.

ADR follows a four-stage process, each stage anchored by principles that capture the essence of the step.

Problem Formulation: For each work included in the thesis in chapters 4 through 7, the first step was to identify the problem being tackled. These problems were, where possible, framed as instances of a broader class of problems, such as XAI or the integration of AI into sleep healthcare. Each identified problem was formulated with guidance from practitioners, such as sleep technologists and project leaders with extensive experience in data collection and research. The selected problems served as the foundation and guide for developing the software artifacts, ensuring that the solutions contributed to both practical and theoretical knowledge generation.

Building, Intervention, and Evaluation: Every work included in this thesis followed an iterative process of designing, interviewing stakeholders and experts, and evaluating the produced artifacts. Each artifact was modeled to fit its real-world application and continuously refined and iterated with feedback collected from stakeholders and end-users.

Reflection and Learning: An integral part of the research process was reflection, where the lessons learned during the artifact design and evaluation were systematically captured in scientific writing. The emergent nature of the artifacts was recognized, with refinements being driven by anticipated and unanticipated real-world scenarios and edge cases. This reflective process guided adjustments to the research methodology while contributing to developing theoretical insights gained by solving the identified problems.

Formalization of Learning: The knowledge generated through the ADR process was presented in scientific texts, submitted, and accepted to various computer science and medical journals and conferences. The formalized lessons and outcomes of the ADR process were structured and positioned to provide guidance for addressing similar classes of problems for future work while contributing new knowledge to the broader existing body. The research outcomes were articulated to ensure relevance to both an academic and practical audience, bridging the gap between theory and practice.

3.2 Data sources

The works included in this thesis are varied and require data sources from diverse data collections. All data used in the articles in chapters 4 through 7 were pseudo-anonymized before being accessed, as covered further in section 3.3.
Table 3.1 outlines the four publications and the data used for each work.

Table 3.1: Data sources per publications.

Work	Data used
BreathFinder: a Method For Non-Invasive Isolation of Respiratory Cycles	31 overnight PSGs [93]
Exploration of Sleep Events in the Latent Space of Variational Autoencoders on a Breath-by-Breath Basis	100 individual sleep apnea testing studies [94]
An optimized framework for processing multicentric polysomnographic data incorporating expert human oversight	Cohort of 50 participants scored by ten independent sleep technologists to create a consensus scoring [95], [96].
World of ScoreCraft: Massively Multi-Scorer Online Study on the Effects of Including Decision Support System in Sleep Stages	Post-participation study, and PSG dataset where traditional and self-applied PSG were recorded simultaneously [97].

3.3 Ethical considerations

When working toward integrating human-affecting AI, such as medicine, insurance, or education, the social consequences must be considered and adequately addressed. Some AI models have been found to harbor various biases [98]; for AI to truly benefit humanity, these biases must be mitigated in some way or another. One potential mitigation for the inherent biases of AI models is the human oversight factor, which is championed throughout this thesis and is one of the key takeaways of this work. For AI to be responsibly utilized in healthcare, measures must be taken to make the decisions made by the AI transparent and explainable. Chapter 5 demonstrates how the contextualization of the data AI is trained and operates on can improve transparency and explainability.

The sensitive nature of healthcare data necessitates stringent privacy measures. In this research, data anonymization techniques were employed to protect the identity of the patients used for validating the BreathFinder algorithm in chapter 4, for training the AI model in chapter 5, evaluation of the effects of AI in chapter 6, as well as the PSG used in the ScoreCraft study in chapter 7 The integration of AI does not exclusively affect the patients. The effects of AI on sleep technologists have the potential to be a positive effector in terms of scoring speed or accuracy, but can also pose behavioral risks such as overreliance on AI tools or de-skilling of sleep technologists. As AI models are autonomous programs that map one input to an output, the models themselves can never be held accountable for any mistakes or harm caused by said mistakes. Therefore, accountability must reside with the medical professionals who utilize the output of the AI models. The work done in this thesis supports this, as chapters 6 and 7 emphasize the need for human-in-the-loop systems rather than for AI to replace the human expert.

As stated before, while AI promises to enhance diagnostic accuracy and reduce clinical workloads, it also raises concerns about the loss of human touch in healthcare and the risks of over-automation in a field where humanity and empathy are paramount. For AI developers, the goal should be to ensure fairness and transparency in the outputs of AI models. In general, those who seek to integrate AI into healthcare professionals' workflows should make it their mission to develop systems and workflows that maintain the human as the final decision maker. Acknowledging and addressing these ethical considerations ensures that AI tools in sleep research and medicine are developed and deployed to benefit all stakeholders without compromising ethical standards.

3.4 Researcher's role

Per Reykjavik University regulations on authorship contribution statements, table 3.2 outlines the publications in chapters 4 through 7 and states the contribu-

tion status of the Ph.D. student (Benedikt Holm, in this case) in various tasks and phases of the publications. The contribution levels, in order of work magnitude: Major Effort (ME) denotes that the Ph.D. student contributed major effort towards the tasks related to the indicated column. Equal Effort (EE) indicates that the Ph.D. student contributed an equal effort to the indicated column, for example, if the drafting process was shared or if work was otherwise divided between co-authors. Contributing Effort (CE) is assigned when the Ph.D. student contributed some effort towards the indicated column, but another author delivered the main effort. Learning Effort (LE) is used when the Ph.D. exerts an effort of a learning character, for instance, by assisting with the data collection or the analysis; however, it does not significantly contribute to the indicated column.

Table 3.2: Declaration of author contribution.

Paper name	Idea	Related work and literature	Data gathering	Research Design	Artifact Design	Analysis and synthesis	Draft	Administration
BreathFinder: a Method For Non-Invasive Isolation of Respiratory Cycles	ME	ME	ME	ME	ME	ME	ME	ME
Exploration of Sleep Events in the Latent Space of Variational Autoencoders on a Breath-by-Breath Basis	ME	ME	ME	ME	ME	ME	ME	ME
An optimized framework for processing multicentric polysomnographic data incorporating expert human oversight	ME	EE	EE	ME	ME	EE	EE	EE
World of ScoreCraft: Massively Multi-Scorer Online Study on the Effects of Including Decision Support System in Sleep Stages	ME	ME	ME	ME	ME	ME	ME	ME

Chapter 4

BreathFinder: a Method For Non-Invasive Isolation of Respiratory Cycles

4.1 Introduction

At present, in order to be able to correctly diagnose a sleep disorder, an expert sleep scorer must manually review (score) a Polysomnography (PSG) which is an overnight collection of various physiological signals from a patient suffering from a suspected sleep disorder. This type of study is performed either in a controlled hospital environment or in a home setting, each with their advantages and disadvantages [99]. A wide range of signals is currently being collected, including respiratory inductance plethysmography (RIP), oxygen saturation (SpO2), nasal airflow, electroencephalography (EEG), electromyograms (EMG), electrocardiography (ECG), audio, and others [8]. The sleep scorer must annotate the sleep stages and other events of interest, which include respiratory events (apneas, hypopneas), oxygen desaturations, body movements, and respiratory event related arousals (brief waking periods due to breathing interruptions). These annotations are then used to determine the diagnosis and recommend a treatment.

For historical reasons, most automated scoring of PSG data uses fixed-length epochs following the methodology adopted for manual scoring. An alternative approach, adaptive segmentation, which is based on segments of variable length depending on the signal [27] has to the best of our knowledge been confined to research on brain activity during sleep [100], with some success in sleep staging [101], [102].

A limited amount of literature exists on using adaptive segmentation to identify individual breaths, or respiratory cycle isolation (RCI), with existing methods

mainly based on statistical analysis of signals such as peak and valley detection in the airflow, thoracic or abdominal RIP signals, and feature extraction and modeling, which are mostly derived from the audio signal recorded during the study. [37]–[45]. One problem with some approaches is that formal validation of the algorithm is often not provided on patient data. Issues with equipment, wide variations in patient behavior, and noisy environments such as partner breathing or background noise, may consequently not have been adequately explored.

Rosenwein et al. introduced a breath detection algorithm based on a random forest approach [40]. They derived 351 features from audio recordings and trained the model to detect inspirations and exhalations, reporting an 87% and 76% accuracy in predicting inspiration and expiration events, respectively. Yaha & Faezipour trained a support vector machine to detect respiratory phases using audio from a microphone placed in front of the participant's nose [41]. They report a 95% accuracy but acknowledge that the results depend on a number of factors of the microphone's placement and quality. Palaniappan et al. developed a neural network solution for classifying respiratory phases using respiratory sounds [46]. While they reported very good overall performance for their model, they did not disclose the accuracy of the model for the different breath phases. Hsiao et al. designed an attention-based autoencoder to perform respiratory segmentation on the audio signal [45]. They report a 91% accuracy in detecting the respiratory phases. A similar method applying adaptive segmentation using a variable window size was also proposed by Lalouani et al. with the AUDAS system [47], which is capable of segmenting the audio signal into individual breath phases as well as classifying chronic obtrusive pulmonary disease with a reported 92% accuracy in the detection of breaths vs. non-breaths with their model. Hult et al. proposed a bioacoustic method that can accurately time respiration from tracheal sounds, using a summation method over the frequency domain of the audio signal. They evaluate their algorithm on 2074 respiratory cycles from two groups of participants, one being recorded in a quiet environment and another with acoustic disturbances from surrounding activities. They report detecting respiratory phases with 99% accuracy on participants in the group with the quieter environment, and approximately 90% accuracy in the group recorded with more noisy conditions [44]. Alshaer et al. present a method for segmenting respiratory audio from a cardioid microphone but do not explicitly state performance metrics [48]. Moyles & Erlandson proposed a non-parametric statistical approach to RCI, based on detecting changes in the trend of the airflow signal, but do not provide any validation of their algorithm [37]. Korten & Haddad presented a pattern recognition algorithm that detects respiratory events in a barometric pressure signal [49], they claim the difference in mean values for inspiratory time (T_i), expiratory time (T_e), and total respiratory cycle time (T_tot) between the manually calculated values and the automatically detected values using the pattern recognition algorithm is very small (<6%), but do not explicitly state performance in terms of detections. Lopez-Meyer et al. pre-

sented an RCI algorithm based on peak and valley detection on RIP signals to determine the beginning and end of a breath segment, reporting 96% precision when detecting breath cycles for participants during rest[39]. A Python library, RespInPeace, for RIP belt analysis is also available, which uses a peak and valley location algorithm to find respiratory cycles during a conversation. but appears to have no published validation [42].

Although the existing literature presents diverse methods for RCI algorithms, there is no consensus yet on which signals to base the segmentation on, validation of methods is limited and whether the segmentation should be done on a respiratory phase basis, or on a respiratory cycle basis is not always clear, with the task of RCI sometimes referred to as breath segmentation [39], [103], or breath cycle segmentation [46]. The problem of performing RCI on PSG data in relation to sleep and respiratory events does not appear to have been deeply studied.

In the rest of this paper, we will present and evaluate BreathFinder, a novel algorithm that locates individual respiratory cycles within breathing signals collected from a PSG using signal processing and statistical methods. The aim of this research is to enhance the analysis of sleep data on an individual breath level. We evaluate our method on a real-life dataset of over eight thousand individual breaths.

The main contributions of this paper are:

- A novel algorithm for performing RCI.
- New methods for evaluating the performance of algorithms designed to locate events in signals.

4.2 Materials and methods

The common definition of a respiratory cycle in the literature splits a single cycle into 4 distinct phases; inspiratory, inspiratory pause, expiratory, and expiratory pause [43]. In this work, a single cycle in the respiratory system is defined as starting with an inhalation and ending just after the following exhalation, with the terms ‘breath’ and ‘respiratory cycle’ considered synonyms. This definition ignores the inspiratory pause phase, and interprets the expiratory pause phase as a pause between two individual breaths belonging to neither. This definition also explicitly defines that by its definition of breath, no two breaths can occupy the same moment in time. The different phases of the respiratory cycle as defined in this work are visualised in Figure 4.1.

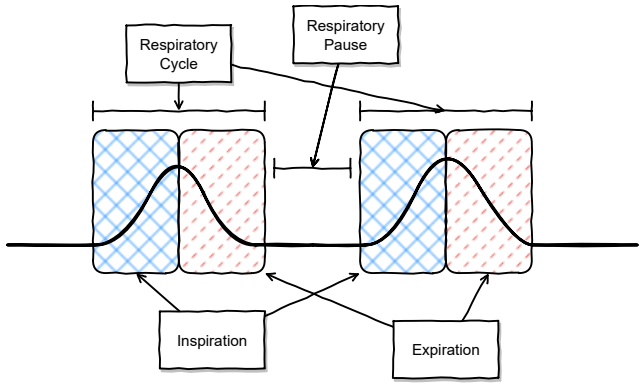


Figure 4.1: Phases of the respiratory cycle.

We used breathing signals from the PSG to detect respiratory events, in particular the airflow and RIP signals. The airflow signal measures nasal respiration, and is most commonly measured with a pressure transducer attached to a nasal cannula[8]. RIP signals are measured via two belts which stretch around the thorax and abdomen to measure changes in inductance caused by the movement of the body part they are placed around. RIP belts are normally used to detect respiratory events in conjunction with the nasal cannula and to estimate respiratory effort[8].

When performing RCI, a decision must be made on the signal source which is most appropriate for this purpose. The two main factors in this decision are the error rate of the signals and any potential impacts of external factors such as background noise.

In practice the nasal cannula has several logistical issues: the sensor can get loose, affecting the measured airflow, or the participant can start mouth breathing, bypassing the sensor completely. Since multiple studies also show that the nasal airflow signal exhibits poor quality in an estimated 10% of cases [99], [104], we deemed it inappropriate for this study. The audio signal was also eliminated, even though some studies show the signal is not as prone to error as the other signals [104], because the signal may contain many different acoustic events such as snoring, movement-related artefacts, or various background noises which complicate the task of pure breath detection[44]. RIP belts have the advantage that they are not susceptible to ambient noises, nor the bypass problem that the airflow signal may encounter. Of the two RIP belts, since the thoracic RIP signal captures the action of the chest-wall muscles more closely than the abdominal signal, we chose the thoracic RIP signal as the basis for our analysis.

4.2.1 The BreathFinder algorithm

A flowchart of the proposed RCI algorithm is presented in Figure 4.2. The algorithm takes a thoracic RIP signal as a parameter, along with a sampling frequency *fs*. The output is a list of individual respiratory cycles, each consisting of the onset, i.e. the start of the respiratory cycle in seconds since the signal start, and the duration of the respiratory cycle in seconds. The algorithm works on the principle of segmenting the signal into windows $w[n]$, with arbitrary onset n in the signal, and then searching for a single respiratory cycle in the selected window. The algorithm first calculates the autocorrelation function for $w[n]$ in order to estimate the lengths l of all potential breaths in the window. It then uses a probability model to discard breath length candidates that are considered too unlikely, either because the length is too long or too short. Then, for each remaining breath length l, the algorithm creates a template waveform of that length, which it correlates with the signal window to find where in the window the breath onset is most likely to be. After the window is analysed, the algorithm advances the window further in the signal, and repeats the process. The analysis windows overlap in order to allow the algorithm to analyse every breath multiple times.

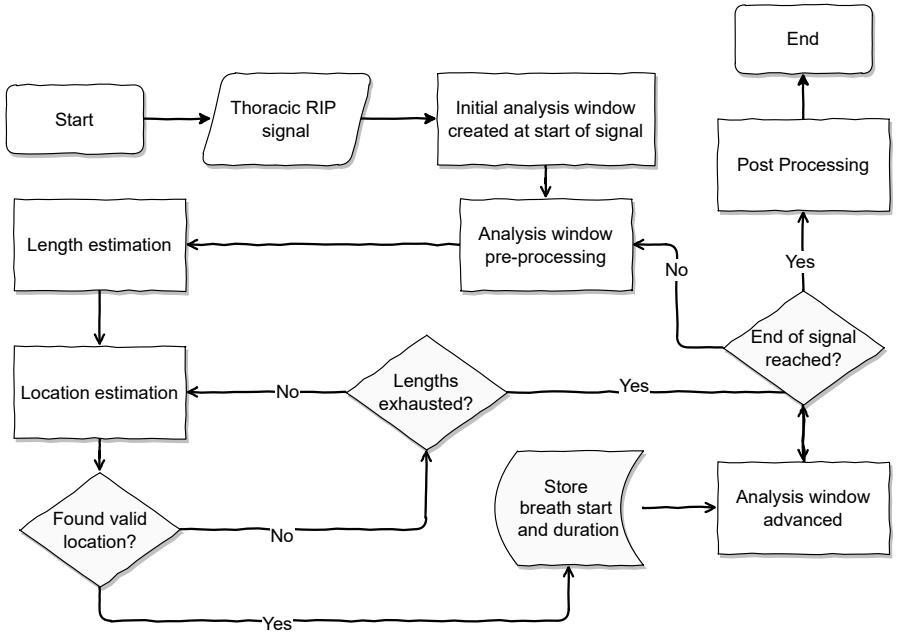


Figure 4.2: Respiratory Cycle Isolation algorithm flowchart.

Signal pre-processing

The preprocessing of the RIP signal is two-fold. First, the entire signal is smoothed, using a Savitzky-Golay filter, which fits a polynomial function to smooth the data points [105]. Here, we used a filter employed with a third order polynomial and a window size of 51. The filter parameters were tuned beforehand via experimentation to ensure that the smoothing had a minimal effect on the overall shape of the RIP signal while still eliminating some of the finer noise. Then, each individual $w[n]$ was corrected for skew. This was achieved by fitting a linear function to the signal in $w[n]$, and adjusting the function so that the y-intercept was 0.0. The function was then subtracted from each sample of the signal. This procedure removes large-scale skew from the signal window, but leaves the general shape of the signal intact. The procedure also had a positive effect on the template waveform fitting procedure, making it less likely to produce incorrect results due to skew. The result of this pre-processing step was that the cleaned thoracic RIP signal was ready to be used to estimate breath lengths.

Main algorithm body

In the first step, the algorithm takes an analysis window $w[n]$, containing a cleaned signal and estimated its periodicity T using the autocorrelation function. The principle of the autocorrelation function (ACF) is to shift the signal forwards in time by k and to compare it to itself. When $k = 0$, the signal correlates perfectly with itself, but as k increases, the correlation decreases. The formula for autocorrelation of a signal x is:

$$\mathrm{ACF}(x)[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[k] \cdot x[n+k], \tag{4.1}$$

where N is the length of x, and k the shift. For periodic signals, when $k = \frac{T}{2}$, the value of the auto correlation is low, as the signal is being compared to itself when it is in asynchony. As k approaches T, the correlation value increases, as the first period lines up with the following period. Thus the peaks of $ACF(w[n])$ can be used to estimate the periodicity of a signal [106]. In the next step, the peaks in $ACF(w[n])$ were found using a peak-finding algorithm. Since the analysis window length was more than twice the mean breath length, $w[n]$ is likely to contain at least two breath cycles, and thus multiple peaks. To address the possibility of false alarm peaks produced by this approach, the algorithm models the breath length probabilities with a normal distribution. The parameters μ and σ for the normal distribution were calculated as the mean and standard deviation of the length of breaths within Evaluation-Subset-A and B.

The modelled probability distribution can be seen in Figure 4.3 along with a density histogram of the breath lengths from the sets of 8782 manually annotated

intervals.

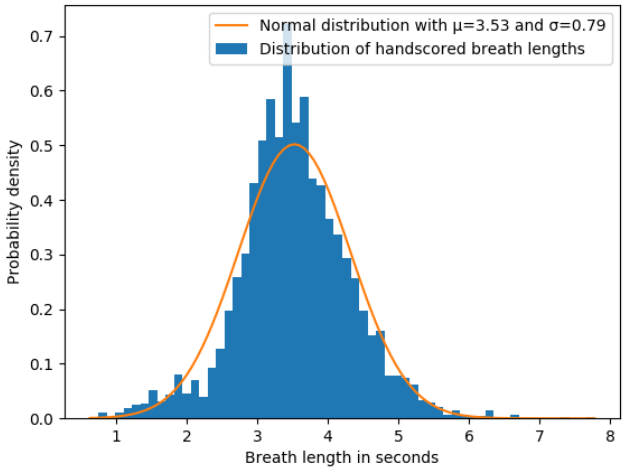


Figure 4.3: Reference breath length histogram with model normal distribution.

Using this probability distribution, the algorithm can rank the breath length candidates, ensuring that it considers the most probable breath length first, thus saving on computing time. Additionally, any breath length candidate whose length probability is less than three standard deviations from the mean is discarded as being too unlikely. Practically speaking, this means that any breath shorter than approximately 1 second or longer than 6 seconds was discarded.
In the next step, for every remaining breath length candidate, a discrete sine template waveform is generated, using the following formula:

$$sin[n] = sin(\frac{n * 2 * \pi}{l} + \theta), \tag{4.2}$$

where l is the length of the candidate in seconds, and θ is an offset that can be set to 1.5 π to shift the waveform so that it starts at -1, ends at -1, and has a peak in the middle. To find where a given template waveform fits most closely to the RIP signal window, the algorithm compares it to the RIP signal using the Pearson correlation coefficient (ρ) at each point on the RIP signal. The formula for calculating ρ for a pair of signals is:

$$\rho(\text{w[n]}, \text{sin[l]}) = \frac{\text{cov(w[n]}, \text{sin[l])}}{\sigma_{w[n]}\sigma_{sin[l]}} \tag{4.3}$$

where $cov(w[n], sin[l])$ is the covariance of the window and template waveform , and the covariance can be calculated as:

$$\text{cov(w[n], sin[l])} = \sum_{i=1}^{N} \frac{(w[n][i] - \bar{w[n]})(sin[l][i] - \bar{sin[l]})}{N} \tag{4.4}$$

where N is the length of w[n] and sin[l] which must be equal, and $\bar{w[n]}$ and $\bar{sin}$ are the means of the respective signals. The sign of ρ describes whether the signals are positively or negatively correlated, and its value describes how strong the correlation is. A ρ value of 0.0 means that the signals are not correlated, a value of 1.0 indicates a positive correlation and a value of -1.0 means that the variables are perfectly inversely correlated. The algorithm treats the RIP signal as one variable and the template waveform signal as another, and calculates the correlation of the template waveform over the entire window. The correlation of the template waveform and the RIP signal produces a third signal, whose peaks represent possible onsets of the target breath.

Since the template waveform is an approximation of the shape of a breath, ρ is not expected to reach 1.0. However, the correlation still provides information about the validity of the breath onset. The algorithm discards any breath onset candidate whose ρ is less than 0.75. The ρ elimination criterion was chosen via experimentation to eliminate as many inaccurate guesses as possible, while still not being so strict as to eliminate legitimate guesses on noisy data, at approximately 0.5ρ below the stable elimination criterion (see Fig. 4.6c). If this elimination step filters out all breath onset candidates, the algorithm repeats the template waveform fitting process with another breath length candidate. If the algorithm processes all breath length candidates and no breath is found in the current signal window, then the algorithm moves on to the next window. If the correlation is above the threshold, the algorithm adds the onset and the duration to a list of breaths and moves the window onset to the end of the detected breath. This process is repeated until the signal is fully analysed.

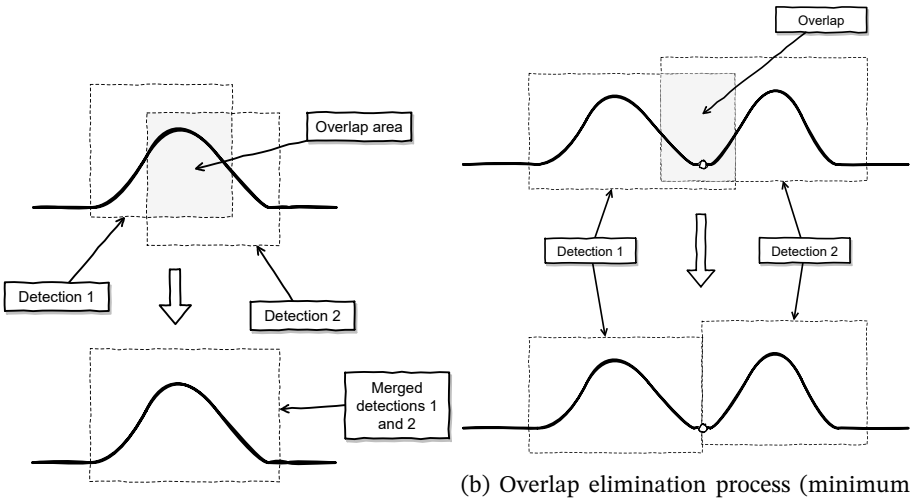
4.2.1.1 Breath placement post-processing

As the sliding windows overlap, the algorithm has a tendency to rediscover breaths. To solve this problem, the i^{th} breath is compared to the $i+1^{th}$ breath. If the overlap of the breaths spans the majority of the total length, the breaths are considered a double detection, and therefore the detections are merged. The process of merging two breath detections involves replacing them with a single detection which covers the area that both previous detections covered. The percentage overlap calculation for a pair of time spans is:

$$O_w(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4.5}$$

where $|A|$ and $|B|$ are the lengths of time spans A and B respectively, and $|A \cap B|$ is the overlap area of detections A and B. If the breaths do not overlap at all, the value produced by this function is negative, and in the case of perfect overlap, the overlap value is 1.0. For this reason, the function is clamped above 0.0. The detection merging procedure is visualised in Figure 4.4a.

Due to the 80% overlap required to merge breaths, the breaths can still overlap by up to 20%. By definition, a breath cannot overlap with another breath, so for each breath, the i^{th} breath is compared to the $i+1^{th}$ breath. If they still overlap, the end of the i^{th} and start of the $i+1^{th}$ breath are moved to the time of the minimum value of the RIP signal within the overlapping region. This process is visualised in Figure 4.4b.



(a) Detection merging procedure.

(b) Overlap elimination process (minimum of the RIP signal marked by a white circle).

Figure 4.4: Key processes in the algorithm. (a) Detection merging procedure. (b) Overlap elimination process.

The end result of this post-processing process is that there is no overlapping pair of detections, satisfying the constraint that no two breaths can share a moment in time. When run on Evaluation-Subset-A and Evaluation-Subset-B, the post-processing step removed 2.05% of detections on average from each interval.

Algorithm evaluation

As the algorithm's task is to mark an individual detection event anywhere on a signal, an obvious problem presents itself when comparing detections to a ground truth. If the algorithm produces a false positive, splits a single breath into two

or more breaths, or any case in which an extra detection is inserted, a misalignment between the list of detections and annotations is created where the detections placed after the false positive have an index that corresponds to the index of a later annotation than it should. The error compounds after each false positive. The same misalignment error is created when the algorithm produces a false negative, except the misalignment is now reversed, i.e., each detection after the false negative has an index corresponding to the index of an earlier annotation than it should. As with the previous case, the misalignment error compounds after each false negative.

Due to the possibility of misalignment errors, it is not possible to naively compare the list of detections and annotations, and an extra step has to be performed to match detections to their corresponding annotations. We solved this alignment problem by using a matrix containing the percentage overlap of all available pairs of detections and annotations calculated using eq. 4.5. This matrix is referred to as the overlap matrix and simplifies the process of finding which detection corresponds to which annotation, whether or not a given detection is a false positive or not, and whether a given annotation corresponds to a detection or is a false negative. Given an overlap matrix A of a list of detections X and a list of annotations Y, the overlap of any $X[i]$ and $Y[j]$ can be accessed in $A[i, j]$.

Using an overlap matrix, a detection corresponding to any annotation could be found by locating the index of the maximum overlap value in the overlap matrix column for that annotation. To be counted as a correct detection for a given annotation, a detection must have a weighted overlap value of over 80% with that annotation. If an annotation had no value above that threshold in its column in the overlap matrix, the breath was counted as having been missed by the algorithm (false negative). Similarly, if a detection had no value above the threshold in its row in the overlap matrix, the detection was counted as a false positive. Due to the restriction that detections may not overlap, it was impossible for two detections to correspond to the same annotation. This paper uses the precision (ratio of true positives to true positives and false positives), recall (ratio of true positives to true positives and false negatives), and F1 score metrics to estimate the accuracy of the algorithm.

In addition to the precision, recall, and F1, additional statistics were collected on the placement error of the detections that were counted as correct. Those include the length of detections versus the length of the annotations. The start and end error was calculated using the following two formulae:

$$\text{start error} = s - \hat{s} \qquad \text{end error} = e - \hat{e} \tag{4.6}$$

where s is the annotation start, e is the annotated breath end, $\hat{s}$ is the predicted breath start, and $\hat{e}$ is the predicted breath end.

The algorithm has four main parameters; the analysis window length, the overlap threshold, the correlation cut-off for the sine fitting procedure, and the probability threshold for the filtering process. To gauge the effect these variables have on the algorithms performance, the evaluation was repeated for a range of values for each variable.

4.2.2 Data description

An extensive evaluation of the correctness of the algorithms output is required, both during normal breathing, and other conditions that may arise during sleep. The dataset used for validation contained 31 overnight PSGs from a population of people diagnosed with obstructive sleep apnea, as well as people with no known sleep issues (VSN-14-080). Of the participants, 13 were female and 18 were male. The mean age of the participants was 47.1 years, in the range of 20 - 69 years. The mean body-mass index (BMI) was 29.9 kg/m^2 in the range of 21.6 - 49.3 kg/m^2. The mean apnea-hypopnea index (AHI) was $9.3h^{-1}$ in the range of 0.0 to 34.8 h^{-1}. Due either to signal failure in the RIP signal, or errors in exporting the recordings from the proprietary NOX format to the standard European data format (EDF), 5 recordings had to be discarded. Each PSG included all standard signals, including EEG, EOG, EMG, ECG, airflow recorded with a nasal cannula, thorax and abdomen RIP belts, pulse oxymetry (SpO2), and an audio signal. The RIP belts in the dataset were recorded with a 25 *Hz* sampling frequency. Additionally, esophageal pressure was recorded with a nose fed catheter [93].

The algorithm was evaluated against 39 variable-length manually annotated evaluation intervals, which further split into 2 evaluation subsets. The first set, referred to as Evaluation-Subset-A was selected to specifically contain various sleep disordered breathing (SDB) events, as well as different sleep stages. Evaluation-Subset-A contained 14 variable length intervals with the mean length of 16 minutes, in the range of 1.5 to 37.5 minutes, with the cumulative length of 225.65 seconds (3.6 hours). The SDB events in Evaluation-Subset-A included obstructive apneas, hypopneas, and increases in respiratory effort without apnea or hypopnea. Further events included in Evaluation-Subset-A were sleep stages, movements, oxygen desaturations and snoring.

These intervals, however, were only selected from one participant in the dataset, and are not representative of the general public. To address that issue, a second evaluation subset was defined, referred to as Evaluation-Subset-B, consisting of a collection of 10 minute intervals from the remaining 25 valid PSGs in the dataset. These intervals were selected randomly from each recording in order to avoid cherry-picking favourable intervals. The random selection was done blindly, aside from being restricted from 1 hour after the recording starts to 1 hour before the recording ends. This was done to reduce the probability of including either the participant settling down to sleep, or moving around as they wake up.

The locations of individual breaths in both evaluation subsets were then manually marked using a custom-made scoring tool programmed in Python. The manual breath annotations represented the ground truth. Of the total 39 intervals in Evaluation-Subset-A and Evaluation-Subset-B, one was found to be incorrectly manually annotated and was discarded. The algorithm was therefore evaluated on 7.3 hours of manually annotated data over 38 intervals, containing 8782 individual breaths from 26 participants.

4.3 Results

The algorithm was evaluated on two sets of manually annotated intervals, the first set containing a relatively high amount of SDB events, and the second set being sampled from a population of 25 participants. Fig. 4.5 shows the format of how a breath is detected by the algorithm. The results of the performance evaluation are

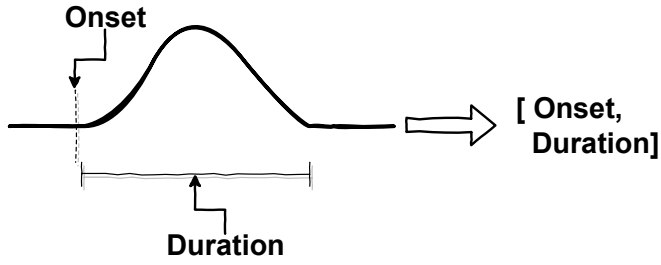


Figure 4.5: Visualisation of an example detection.

summarised in table 4.1, and the placement errors are shown in table 4.2.

Table 4.1: Evaluation results of the RCI algorithm.

Evaluation subset	Mean precision	Mean recall	Mean F1
A	0.94	0.94	0.94
B	0.93	0.95	0.94

The algorithm achieved, on average, 0.94 precision for Evaluation-Subset-A and 0.93 for Evaluation-Subset-B. This means that only 6%, and 7% of detections were classified as false positives for Evaluation-Subset-A and Evaluation-Subset-B respectively. The recall for Evaluation-Subset-A and Evaluation-Subset-B was 0.94 and 0.95, respectively, meaning that the algorithm only missed 6% of breaths in Evaluation-Subset-A and 5% of breaths in Evaluation-Subset-B. Two intervals in

Evaluation-Subset-B had noticeably worse results, with F1 scores of 0.79 and 0.81. Upon visual inspection the errors seemed mainly to be due to incorrect manual annotations, and noise in the signal during those intervals. Omitting these two intervals increased the mean F1 of the algorithm to 0.95 for Evaluation-Subset-B. The algorithm performed noticeably worse for one interval in Evaluation-Subset-A than for the others, its precision being 0.76, recall being 0.963, making for an F1 score of 0.854. This was due to the interval being relatively short, and the beginning of the signal being dominated by a movement event, causing the algorithm to misclassify the movement as breaths. On the other hand, the algorithm achieved perfect precision for two intervals in Evaluation-Subset-A and one in Evaluation-Subset-B, all of which contained no SDB events, and only stable breathing. The recall was slightly more stable than the precision for both sets, with the standard deviation being 0.054 for the recall, and 0.058 for the precision.

Table 4.2: Placement errors of the RCI algorithm.

Evaluation subset	Annotation mean breath length (seconds)	Detection mean breath length (seconds)	Mean abs. start error (seconds)	Mean abs. end error (seconds)
A	2.57	2.76	0.16	0.24
B.	3.56	3.88	0.23	0.30

The mean start error for both Evaluation-Subset-A and Evaluation-Subset-B was approximately 6.4% of the mean breath length. The mean end error for both sets was greater, being 10% and 8.4% of the mean breath length for Evaluation-Subset-A and Evaluation-Subset-B respectively. When visually inspected, the alignment of the detections and the thoracic RIP signal was high for both Evaluation-Subset-A and B.

4.3.1 Sensitivity analysis results

The analysis window length is the length in seconds of the window that the algorithm uses to search for breaths in at each step.

The results of the sensitivity estimation can be seen in Figure 4.6a, and shows that both precision and recall rise sharply as the window length reaches approximately 6 seconds, and plateaus at approximately 8 seconds. The reason for the sharp rise in performance between 2-6 seconds is most likely that the window can not reliably fit two cycles of the respiratory cycle until the window becomes longer than twice the average length of breath in the dataset.

The overlap percentage is the amount of the previous window included in the next window as the analysis window advances. The effect of this parameter is

shown in Figure 4.6b, which suggests that the algorithm performs noticeably badly only in terms of recall when the overlap percentage is around 0. This can be explained by the algorithm missing breaths as the window skips either entirely, or partly over them. The precision seems largely unaffected, which indicates that the number of false positives drops proportionally with the number of true positives.

The correlation threshold dictates how much a breath candidate must resemble a model breath in terms of its Pearson correlation (Figure 4.6c). It is meant to filter out waveforms that may only superficially resemble breaths, however, still forming peaks in the sine-correlation function.

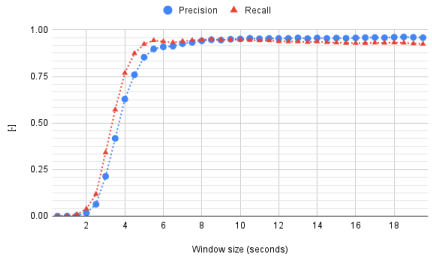
As the correlation threshold increases, the precision improves. This can be interpreted as the criterion for what ”looks like a breath” becoming stricter, thus eliminating more false negatives. The recall seems unaffected by this criterion until the threshold reaches approximately 0.8, at which point it sharply drops. This is to be expected, since the template waveform is only an estimation of the general shape of a breath in the signal, and thus the correlation with the signal is not expected to be perfect.

The probability threshold parameter is used to discard breaths that are considered too improbable. To estimate the effect of this variable on the performance of the algorithm, the evaluation was repeated for a range of probability thresholds.

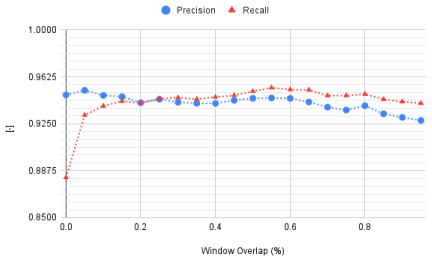
As Figure 4.6d indicates, the absence of this filtering step seems to have little effect. The precision is least affected by the probability threshold, while the recall drops sharply off as the threshold increases. This is reasonable since as the threshold increases more legitimate breaths are discarded, thus negatively affecting the recall until the threshold reaches approximately 0.55, at which point all breaths are discarded. The stability of the precision suggests that the rate of false positives drops proportionally to the rate of true positives as this parameter approaches 0.5. The reason for the falloff of both the precision and recall at 0.5 is that the maximum possible value of the probability estimator is 0.5, any value above 0.5 will therefore cause the filtering process to discard all detections.

4.4 Discussion

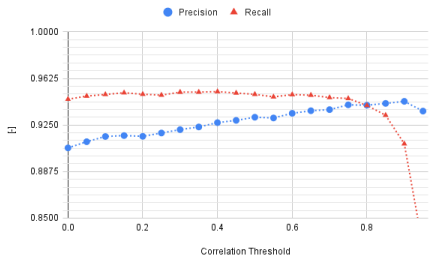
This paper presents a novel algorithm designed to perform RCI on the thoracic RIP signal, based on signal processing and statistical methods. The algorithm achieved an F1 score of 0.94 when detecting breaths during sleep over multiple nights and including SDB events, that is 94% of breaths were classified correctly, with 6% false negatives. Of the detections made by the algorithm, approximately 95% are correctly placed breaths, with only 5% being false positives. This accuracy is superior [40], or comparable [39], [43] with previous work, however, we note that comparison to some prior work is problematic since there is no standardized method of evaluating RCI algorithms, and thus different works approach the task of evalua-



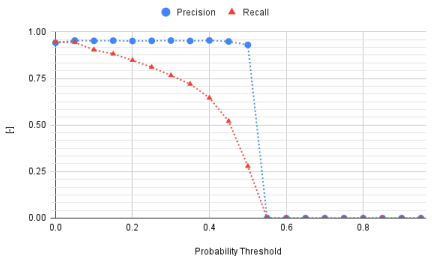
(a) Algorithm Recall and Precision sensitivity to window length.



(b) Algorithm Recall and Precision sensitivity to overlap percentage.



(c) Algorithm Recall and Precision sensitivity to correlation threshold.



(d) Algorithm Recall and Precision sensitivity to probability threshold.

Figure 4.6: Sensitivity analysis of algorithm Recall and Precision to various parameters. (a) Window length, (b) Overlap percentage, (c) Correlation threshold, and (d) Probability threshold.

tion differently, making comparisons difficult, if not at times impossible [37], [42], [46]. Comparison between algorithms can be seen in Tab 4.3

Currently, the algorithm is only evaluated on RIP signals collected with a 25Hz sampling frequency. The algorithm is designed to be independent of the sampling frequency of the signal, but requires a similarly rigorous evaluation at other sampling frequencies. In the validation data used in this paper, the algorithm is validated on intervals containing significant movement, respiratory events, as well as various sleep stages.

The evaluation found that the detection rate was not meaningfully influenced by respiratory events or physiology, however, the most impactful factor in terms of detection rate seemed to be artefacts. On the other hand, artefacts such as movement or signal failure only cause the algorithm to produce errors where the artefacts occur, and cause no errors for future detections, indicating that the algorithm

Table 4.3: Comparison between this work and related work.

Algorithm	Accuracy	Signal	N
BreathFinder (This work)	**94%**	**Thorax RIP**	8782 respiratory cycles
Rosenwein et al. [40]	87%	Audio	188.850 events (inspirations)
Yaha & Faezipour [41]	95%	Audio	128 events
Palaniappan et al. [46]	"Very Good"	Audio	72000 events
Hsiao et al. [45]	91%	Audio	489 recordings of lung sounds from 22 participants (number of events not specified)
Hult et al. [44]	99% - 90%	Audio	1863 respiratory cycles
Alshaer et al. [48]	Not Stated	Audio	-
Moyles & Erlandson [37]	No Validation	Air Flow	-
Lopez-Meyer et al. [39]	96 %	RIP Belts	At least 10 minutes from 4 participants (number of events not specified)
RespInPeace [42]	No validation	RIP Belts	-

can easily recover from a noisy period. When the detection error of the correctly detected breaths is expressed as the mean absolute start and end error, the algorithm tends to produce greater end errors than start errors. The mean end error, however, was less than 8% of a mean breath length, and upon visual inspection, was not discernible to the human eye. The start and end errors of both sets may be partially explained by the fact that the manual annotations did not observe the restriction that only one breath can take place at any moment imposed by this work's definition of the respiratory cycle, effectively introducing small sections of the annotations that at most one detection can overlap with thus artificially negatively impacting the metrics.

Despite the fact that the algorithm is originally designed for use in sleeping individuals, we believe it could be used to research respiration during speech, exercise, emotional response analysis, and other applications provided that the proper

evaluation of the output correctness is performed. The total number of participants used for the evaluation of the algorithm was 25. This is comparable to other literature, where the range of the number of individuals used for testing similar tasks ranges between none reported, 4, 75, and 140 [37], [39], [40], [42], [46]. In future work, the algorithm could be evaluated on a much larger dataset. The algorithm achieved a higher accuracy than the AUDAS algorithm [47], however, the AUDAS algorithm detects individual respiratory phases whereas BreathFinder locates individual respiratory cycles and therefore direct comparison is not appropriate. Similarly to AUDAS, the work done by Hsiao et al. achieves 92% accuracy when detecting inspirations and expirations, but as the task is fundamentally different to this approach, direct comparison is not appropriate [45]. The algorithm is designed to work on the thoracic RIP signal, but in theory should also work on the abdominal RIP signal. However, this requires validation to assess the validity of the results.

Due to the high detection rate of the algorithm and the relatively low rate of false positives, the authors suggest that the proposed algorithm can be reliably used for future research into the nature of respiration during sleep based on RCI-based adaptive segmentation. However, assessment on larger datasets is needed to evaluate the performance of the algorithm when faced with a more diverse range of respiratory events such as central apneas.

4.4.1 Clinical implementation and applications

The implementation of the BreathFinder algorithm has previously demonstrated its utility in other works, particularly in the identification of Obstructive Apneas. It has been successfully applied in detecting obstructive apneas by applying the BreathFinder algorithm on the thoracic RIP signal to find individual breaths, and then performing machine learning on the thoracic, abdomenal, and flow signals during those individual breaths, exhibiting impressive performance with a substantial F1 score of 0.94 in apnea detection tasks, thus corroborating its efficacy and reliability in this context [107]. This makes it an instrumental tool in the diagnosis and management of sleep-related disorders. Moreover, the BreathFinder algorithm has shown versatility by its effective application in an unsupervised machine-learning context. Encoding the thoracic and abdomenal RIP signals, along with the airflow signal from individual breaths facilitated the exploration of the latent feature space, which consequently allowed the identification of significant clusters of breaths. These clusters demonstrated notable common characteristics, including the incidence of obstructive apneas [107]. This exemplifies the algorithm's capacity for contributing to advanced analytical strategies that expose the intricacies of respiratory patterns. It emphasizes the potential for further exploitation of the BreathFinder algorithm in a myriad of applications, includ-

ing advanced diagnostics, predictive modeling, and personalized therapeutic approaches.

4.4.2 Study limitations

The evaluation has the drawbacks that it is only formally done on the thoracic RIP signal, and the evaluation was only done on data from one dataset. The algorithm has furthermore only been evaluated against a RIP signal using 25 as the sampling frequency. Further research is additionally required to specifically evaluate the effects of events such as changes to body posture, RIP artefacts, incorrect RIP placement, RIP belt stability or other deformations in the signal. The purpose of the algorithm is to isolate breaths, rather than to provide any information or statistics on the nature of the breath further than its location in the signal, and any analysis such as obstructive apnea detection or flow measurement is future work made available by this work. Due to the low number of breaths that the BreathFinder algorithm is evaluated on, the statistical significance of the results cannot be assured and thus an assessment of the algorithm on larger datasets is needed to evaluate the algorithm's performance when faced with a larger and more diverse range of respiratory events such as central apneas, and therefore, this work can be viewed as a proof-of-concept study.

4.5 Conclusion

This paper introduces BreathFinder, a novel algorithm designed to find individual breaths in the thoracic RIP signal. The algorithm uses periodicity estimation and sine fitting procedures to pinpoint the locations of individual breaths within a PSG. The algorithm was evaluated on approximately 7.8 hours of manually annotated breathing intervals. The results suggest that the algorithm detects, on average, 94% of breaths correctly, and of the detected breaths, only 4% on average are false positives. The placement error of the correctly detected breaths was generally within acceptable margins, being less than 10% of the mean breath length. The exceptional performance of the algorithm in terms of the evaluation metrics suggests that it is usable for further analysis of sleep data on a breath-by-breath basis. Unlike previous thorax RIP RCI algorithms, BreathFinder is provided as an open-source algorithm, is also validated against a large range of respiratory events, and demonstrates robustness against signal artifacts, also making it the only RCI algorithm evaluated on sleep data known to the authors.

Chapter 5

Exploration of Sleep Events in the Latent Space of Variational Autoencoders on a Breath-by-Breath Basis

5.1 Introduction

Machine learning has opened up new possibilities in analysing data, enabling computers to perform tasks previously thought to be only conducted by humans. Unfortunately, the fields of medicine, and sleep medicine in particular, have lagged behind in the adoption of machine learning, not fully utilising the unrivalled ability of computers to analyse multidimensional data, and detect correlations and patterns most humans would otherwise be incapable of seeing [108]. Thus sleep medicine is still heavily reliant on manual labour which is both time consuming and costly [109].

Sleep, sleep quality and sleeping disorders are measured with an overnight sleep study, which attempts to measure the physiological systems which we lack full understanding of, as evident by the low scorer agreement for sleep stages and different respiratory events [10], [11]. These measurements are represented by multidimensional time series data of various physiological signals, which are then scored by sleep experts in terms of sleep architecture, respiratory events, arousals and more [7]. When considering that the inter scorer reliability when scoring sleep stages is only 82.6% on average [10], and with the agreement on obstructive apneas being 77.1% and hypopneas being 65.4% [11], it is clear that some aspects of these events are not understood well enough to define rules that scorers can easily follow. The fact that the agreement between scorers on these events is as low as it is

suggests that the scoring rules do not adequately describe the signals that they are used to score and are based more on human intuition than the data itself causing the scoring disagreement.

This creates problems when training supervised machine learning models, as the model will be trained to perform like the human scorer, and therefore, the model trained will most likely agree with the scorer it was trained on but less so on scorings made by different scorers. Application of machine learning in medical research is also difficult as the design of models and features often requires domain knowledge [110]. Machine learning has increasingly been used for modelling of the respiratory system, but is not at the place yet where it can take over medical decisions for patients [111]. In recent times, sleep research has seen a surge in computer methods to automate the scoring of sleep stages and events such as apneas or hypopneas but manual review is still required [112].

This problem can be minimized using unsupervised machine learning, where a model is made to learn the data without including any human assumptions through labels about the structure of the data. Unsupervised learning has been suggested as a means of finding previously undiscovered long terms patterns in sleep data [108]. Autoencoders (AE) are unsupervised machine learning models that consist of two parts; 1) an encoder, and 2) a decoder [113]. The encoders task is to transform the high-dimensional input data to a lower dimension, i.e. to map the input to an encoding in a lower-dimensional space often called the latent space. The decoder's task is to reconstruct the original input from the encoding, with the performance of the model being measured by the similarity of the output to the input. A variational autoencoder (VAE) is an autoencoder where the latent space is regularised, as a result vectors close to each other in the latent space have similar decodings [114].

Furthermore, the need for explainable, and fair machine learning approaches has been pointed out, with clinical applications in particular [115], and AE have been described as notoriously contributing to fair representation of data [116].

In this paper, we developed an unsupervised machine learning approach to study respiratory events in sleep studies on a breath-by-breath basis. Our proposed methodology consists of three steps. First we isolate individual breaths from Respiratory Inductance Plethysmography (RIP) belts using the BreathFinder algorithm [117]. Then we train a VAE on the individual breaths to reduce their dimensionality in order to explore the latent space created by the VAE. Finally, we use clustering analysis to detect clusters of breaths' representation in the latent space. The clusters are then analysed in terms of frequency of scored events occurring within the clusters.

The research questions this paper aims to investigate are the following: I. To what extent do events scored by sleep experts cluster together in the VAE despite a completely unsupervised approach? II. What effect does providing the context of respiration to the VAE have on its ability to represent respiratory signals in a

limited latent space? III. How well can the VAE reconstruct the respiratory signals despite a small latent space?

To our knowledge, this approach is novel, and no prior work exists applying these methods on respiratory signals on a breath-by-breath basis. The main contributions of this paper are: Firstly, to the literature on information systems and machine learning, we show that respiratory data can be encoded by a neural network into a relatively small encoding, and reconstructed from the encoding. Secondly, to the literature on ML in healthcare, we show that the latent space of breaths forms regions that can be sampled from and identified by sleep technologists. Thirdly, to the literature on sleep, we contribute with an understanding that breaths that happen during scored events display a tendency to cluster together in the latent space.

The rest of this paper is organised as follows. In the next section we outline related work in the field of machine learning in respiratory, and sleep research. In the Methodology section, we explain how we setup the experiments. Next we present, and then discuss the results before finally concluding the paper.

5.2 Related work

Sleep disordered breathing (SDB) is a class of disorders in which the process of respiration during sleep is interrupted [7]. SDB disorders such as obstructive sleep apnea (OSA) are suggested to be prevalent, having been found in 43.1% of the general population in Iceland [4], as well as estimated to affect one billion people worldwide [5]. [50] designed and trained a combination of a convolutional neural network and a long short-term memory (LSTM) neural network to classify sleep stages using a single frontal electroencephalogram (EEG) channel, they found that classification accuracy decreased with an increase in obstructive sleep apnea syndrome severity. [51] achieved 79.6% accuracy when classifying one second signal segments into normal, obstructive apnea, and hypopnea classes. They used deep convolutional neural networks to automatically learn the features of the airflow, oxygen saturation, and RIP signals. [52] showed that a k-nearest neighbour classifier could achieve up to 91.2% accuracy in classifying one minute signal segments as apneic or non-apneic when using modest features derived from the respiratory inductance signal. This paper used statistics of breaths marked by peak detection, but did not derive breath-specific features. [53] classified respiratory events as apneic/hypopneic versus non apneic/hypopneic on a breath-by-breath basis using a random tree method. They used intuitive respiratory features such as breathing rate and ”duration to last respiratory event” for the classification, and report achieving 86.3% accuracy when classifying breaths as apneic/hypopneic. [54] achieved a 88.9% accuracy on classifying OSA events using a LSTM network, with an average error in apnea-hypopnea index of 3.0.

Outside medical approaches, VAE has been used for tasks such as anomaly detection [55], text classification [56], and recommender systems [57]. In the medical field, VAE and other AE methods have been applied in various ways, mainly for the electrocardiogram (ECG) [58]–[60], but also with some applications for the electromyogram (EMG) and EEG [61]. The application of VAE and other generative machine learning models for biomedical signal analysis is a largely unexplored field. [58] applied a novel class dependent implementation of the VAE for detecting atrial fibrillation, improving the latent space clustering by simultaneously training a classifier on the latent space with the VAE. [62] proposed an adversarial VAE trained to generate and classify respiratory signals, using the latent space to classify baseline shift breathing irregularities, during training. They trained the adversarial VAE on fixed-length signal segments collected by a stereotactic radiosurgery device. Since the is a great amount of uncertainty between scorers, this approach did not seem appropriate for our application.

5.3 Methods

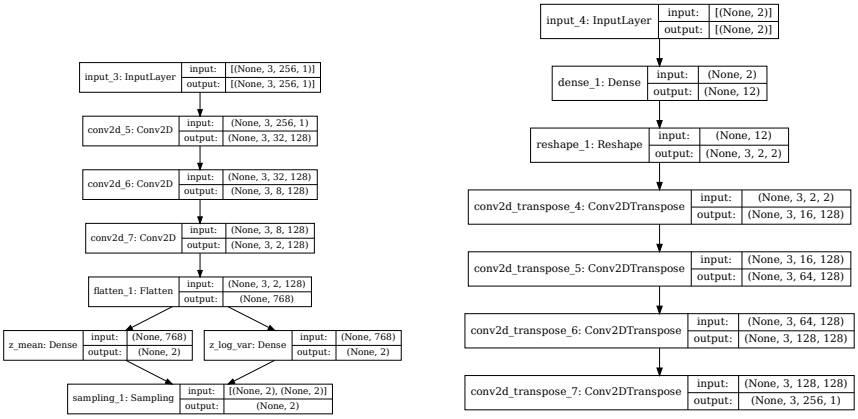
Our proposed methodology consisted of three steps: the segmentation of breaths, VAE, and clustering. Furthermore, for validation, we sampled the reconstructed signals from the clusters.

To locate breaths in the dataset, the BreathFinder respiratory cycle isolation algorithm was used [117]. BreathFinder utilises signal processing methods rooted in statistics to locate individual breaths within the thoracic RIP signal. The extracted breaths were marked as being positive for an event if they overlapped with a manually scored event.

The encoder part component of the variational autoencoder model was constructed using a series of convolutional layers, see Fig. 5.1a, and the decoder was constructed using a series of convolutional transpose layers to restore the original shape of the input from the latent shape. The convolutional layers were chosen, since they are well suited for learning the shapes of the signal, as well as they are suggested to be superior for the task to simple dense layers [62]. The decoder was constructed roughly to mirror the encoder, starting by expanding and reshaping the two-dimensional input, and then passing it to a series of four convolutional transposition layers. The architecture of the decoder can be seen in Fig. 5.1b.

Additionally, as part of the parameter tuning, different layers were experimented with, such as LSTM layers as well as dense layers, with the convolutional VAE coming out on top in terms of reconstruction accuracy. The latent space dimensionality of two was selected, mainly for the ease of visualising the latent space.

The VAE loss function has two terms, the first term is the reconstruction loss, which measures how well the VAE reconstructs the input signal from the latent



(a) Encoder architecture. (b) Decoder architecture.

Figure 5.1: Variational autoencoder architecture components. (a) Encoder architecture. (b) Decoder architecture.

representation, and the Kullback-Leibler divergence [118], which is the term responsible for the regularisation of the latent space, and due to this regularisation term in the loss function, the latent space has the property that vectors adjacent in the latent space will produce similar results when passed to the decoder [114]. For the reconstruction task of the VAE, a reconstruction loss function was chosen with the explainability of the reconstruction error in mind, for this reason, the Mean Squared Error (MSE) was chosen, and is calculated as follows:

$$MSE = \frac{\sum_{i=0}^{|X|}(\hat{x}_i - x_i)^2}{|X|}$$

where x_i is a sample in the input signal X, and $\hat{x}_i$ is the corresponding output sample of the decoder.

The VAE was designed using version 2.4.1 of the TensorFlow [119] machine learning framework, and was written in version 3.9.7 of the Python3 programming language. The chosen optimiser for this work, was the Adam optimiser [120] with a classic learning rate of 0.001. The model weights were initialised with the TensorFlow default random sampling. The VAE was trained on a Linux virtual machine on the Sleep Revolution computational cluster, with a NVIDIA Tesla, V100 32GB, 250W graphics card.

5.3.1 Clustering analysis

To investigate the latent space produced by the VAE, KMeans clustering was applied after passing the testing data through the VAE trained on the training data, see details in Section 5.4.3. We used the standard KMeans implementation in Python's scikit-learn library, which selects initial cluster centroids using sampling based on the data's empirical probability distribution, repeating the process ten times for each setting for number of clusters, to find the optimal result and Euclidean distance as a distance measure. To determine the number of clusters to divide the latent space into, an elbow analysis was performed, with the number of clusters ranging from 4 to 20.

To determine which clusters in the latent space, if any, had a tendency to represent some events rather than others, the encoded breaths were assigned to their respective cluster, and the frequency of scored events was counted for each cluster. The result was a matrix A, where the number of breaths drawn during event j, and assigned to cluster k was accessible in A_{jk}. This matrix was normalised across each row. This was done for two reasons; we were mostly interested in the events that showed an association with one or a few clusters rather than being uniformly spread over the latent space, and we did not want one event that had a relatively large number of associated breaths to be over represented in the exclusivity calculation. To calculate which events showed the greatest amount of association in terms of clusters, the standard deviation of the values in the rows of matrix A were calculated, and the events were ordered by the standard deviation in descending order.

5.4 Experimental setup

5.4.1 Dataset

For the work in this paper, a dataset described by Horne et al. was used which contained 100 individual type three home sleep apnea testing studies, cumulatively spanning 1031.5 hours, with the mean length of recordings being 10.3 hours [94]. The dataset included 71 men, and 29 women, aged between 24 and 68 years old with the mean age of participants in the dataset being 49.4 years and the standard deviation being 10.9 years. The body mass index of participants in the dataset ranged between 17.5 and 52.5 kg/m^2, with the mean and standard deviation being 30.1, and 6.0 kg/m^2 respectively. The data collection was conducted with the consent of the Landspitali Bioethics Committee (22/2014). Written consent was obtained from all participants. The dataset was manually scored for respiratory events, and signal artefacts by a sleep technologist. The types of events and their explanation are outlined in Table 5.1.

Table 5.1: Types of scoring events included in analysis.

Event Type	Description	Count	%
Flow cannula artifact	Signal artifact in the nasal airflow cannula.	8138	2.89
Hypopnea detected in RIP	A hypopnea scored on the RIP signals as nasal cannula was not reliable.	498	0.18
Paradoxical breathing	Paradoxical movement of the RIP signals.	11555	4.11
Position-prone	The breath was drawn in the prone position.	3286	1.17
Pleth pulsewave-drop	Pulse wave signal amplitude drops.	2242	0.8
RIP abdomen artefact	Artefact in the abdominal RIP signal.	843	0.3
SpO2 signal artefact	Artefact in the SpO2 signal	477	0.17
Pulse signal-artefact,	Artefact in the pulse signal	566	0.2
Position-upright	The breath was drawn in the upright position	2965	1.05
Obstructive apnea	The breath was attempted during an obstructive apnea	6898	2.45
Invalid RIP-flow signal	RIP flow signal was invalid, possibly due to an artefact in either RIP signal	883	0.31
Invalid SpO2 value	The value of the SpO2 signal was not valid.	638	0.23
Position-Right	The breath was drawn while the participant laid on their right side	20435	7.26
Movement	The participant was moving while the breath was drawn	16196	5.76
Position-supine	The breath was drawn in the supine position	44412	15.78
Flow limitation	Flow-limited breath	11521	4.09
Oxygen desaturation	A drop in the SpO2 signal, most often associated with a SDB event.	9673	3.44
Position-Left	The breath was drawn while the participant laid on their left side	29098	10.34
Hypopnea detected in Flow	A hypopnea scored on the nasal airflow signal	3323	1.18
Normal Breath	A breath as marked by a sleep expert	72131	25.63
Snorebreath	A breath drawn during a snore	18789	6.68
Snore train	A breath drawn during a snore train	16823	5.98

5.4.2 Preprocessing

In this work we used three respiratory signals namely raw nasal airflow signal (nasal pressure), which measures the pressure generated by the nose during breath-

ing, and the thorax and abdomen RIP signals. We used the BreathFinder algorithm to mark individual breaths, i.e. extract signal segments, from the three signals. A sample of raw respiratory data extracted with BreathFinder can be seen in Fig. 5.2. The total number of breaths across all participants in the dataset was 652,072,

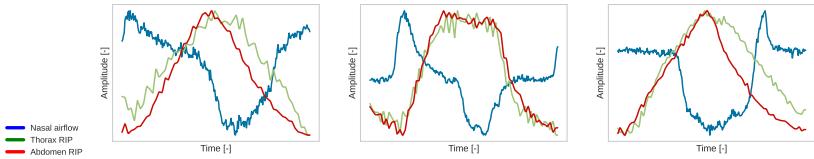


Figure 5.2: Example data of respiratory signals from BreathFinder.

averaging 6520.7 breaths per person. The lengths of each individual breath was normalised to by re-sampling the signals in the breath to 256 samples. The number 256 was chosen due to it being a power of 2, simplifying the structure of the encoder and decoder considerably. Additionally, to eliminate the impact of signal calibration issues on the machine learning process, the amplitude of each signal was normalised such that its amplitude spanned from 0 to 1.

Each breath was marked positive for a type of manually scored event if the breath overlapped at any point with the scored event. Table 5.1 furthermore shows the number of breaths in which each event occurred as well as the percentage of breaths affected by each event.

5.4.3 Data splitting

To train the VAE, the dataset was split into two groups of participants, a training group of 85 participants, and a testing group of 15 participants. The split was done on a participant level rather than on a breath level in order to prevent the VAE from learning to accommodate for the physiology of the individuals it was trained on, and thus achieving better results when reconstructing unseen signals on familiar physiology. The training dataset was composed of 63 males, and 22 females, and the test dataset was composed of eight males, and seven females. One person was recorded twice in the dataset, and thus the second night was discarded from the dataset in order to not add any bias to the training.

The configuration of the VAE was determined with an experimental process, adjusting both number of hidden layers, size of convolutional filters, and changing the learning rate. The parameter tuning process was performed with the goal to optimise the reconstruction loss. During the training process the training dataset was kept constant, to eliminate the possibility of the model ‘getting lucky’ with an optimal combination of training data and model configuration. The testing data was not used in the process of parameter tuning. The best performance in

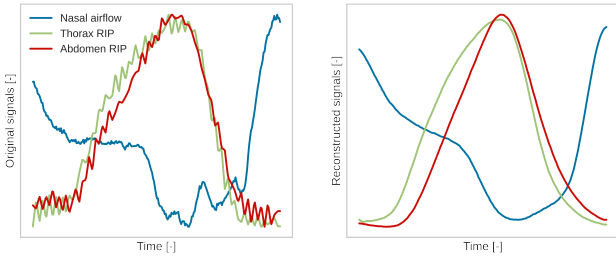


Figure 5.3: Example of the reconstruction of the model.

terms of reconstruction loss was achieved with the encoder was constructed with 3 convolutional layers, followed by a flattening layer, and finally a sampling layer [114].

5.5 Results

5.5.1 Training

The autoencoder was trained on the training data until the training loss converged at approximately 18.9. With 3×256 samples to reconstruct, the loss ended up converging when the reconstruction error was on average 0.024, for each sample in the reconstructed signals. Since the signals were normalised to span between 0.0 and 1.0 during training, this corresponds to approximately 0.15 error on average per sample in the reconstruction signal for the training data. This 0.15 error per sample suggests a considerably bad reconstruction, but as Fig. 5.3 shows, a lot of the higher-frequency signal data has been removed in the reconstruction, but the overall shapes of the respiratory signals are preserved.

5.5.2 Clustering analysis

After training the VAE the testing data was passed through the model and KMeans clustering applied to the encoded two-dimensional data in the latent space. The results from the elbow analysis suggested that the optimal number of clusters was nine as can be seen in Fig. 5.4, the line however, does not show a definitive elbow.

In order to attempt to explain the model, samples were taken from ten different points at the centres of visually identified clusters in the latent space. A sign that the VAE trained correctly is that points close in the latent space had a similar reconstruction, as evident when comparing the similar reconstruction of points c and f, versus points c and d in Fig. 5.6.

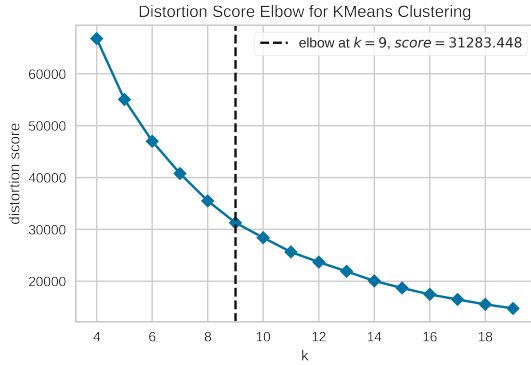


Figure 5.4: Elbow analysis (distortion score is the sum of square errors).

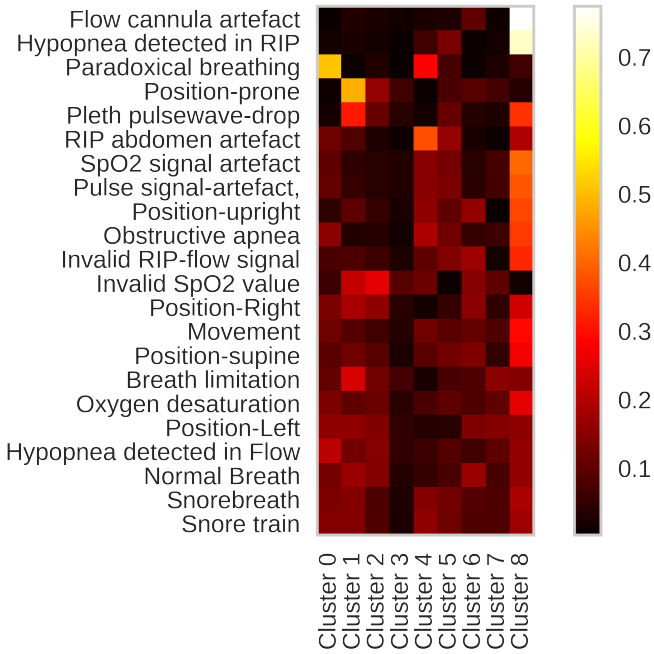


Figure 5.5: Cluster prevalence for scorings.

The KMeans clustering produced the cluster configuration of the latent space which can be seen in Fig. 5.6, and observing reconstructions a, d, e, g, and f in par-

ticular, suggests that the VAE can accurately reconstruct the interplay between the respiratory signals. For example, as the thoracic and abdominal RIP belt signals expand and contract, the airflow signal shows an inspiration, and expiration.

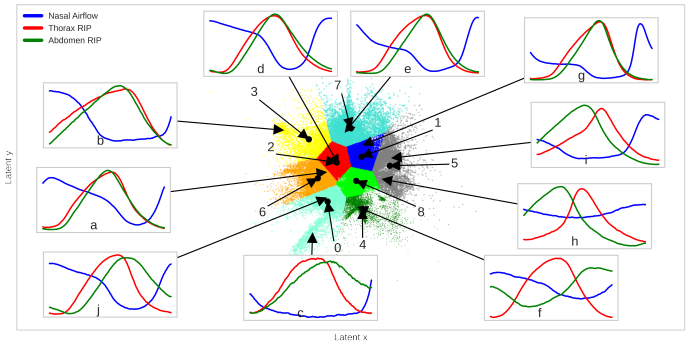


Figure 5.6: Latent space sampling of individual breaths (a-z) with marked cluster centres (1-10).

The cluster association of the manually scored events in the dataset was calculated, and the results are displayed in Fig. 5.5. Different scored events showed a varying level of 'cluster association' and we take a closer look at some of these events and their distribution in the latent space. The highest scoring event in terms of cluster association was the nasal airflow artefact scoring. The nasal artefact breaths tended to be placed overwhelmingly in cluster number eight, and as shown in Fig. 5.7, the airflow artefact breaths clustered closely together in the latent space, indicating that the nasal airflow artefacts were recognised, and learnt by the model.

The second highest rated event in terms of cluster association was the hypopnea events as can be seen in Fig. 5.8, showing a clear association with cluster eight, and a very limited area on the latent space in comparison to the airflow artefacts in Fig. 5.7.

An event of particular interest was the paradoxical breathing often associated with obstructive apneas. As Fig. 5.9 shows, the distribution of paradoxical breathing events is distinctly skewed to lower half in the y-dimension in the latent space, and the x-dimension of the paradoxical breaths forms a slightly bimodal distribution. The breaths drawn in the prone position showed a significant skew towards the upper end of the y-component of the latent space distribution, and slightly to the right of the x-component as can be seen in Fig. 5.10. For comparison, the breaths drawn in the supine position are visualised in Fig. 5.11, which shows a

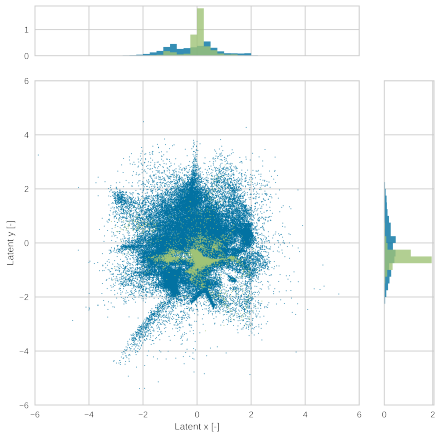


Figure 5.7: Flow artefacts (green) vs. the population distribution (blue) in the latent space.

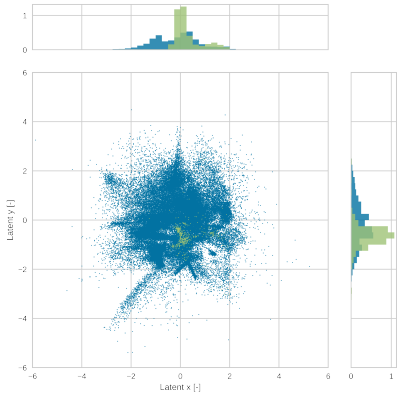


Figure 5.8: Hypopnea events (green) scored on the nasal airflow signal vs. the population distribution (blue) in the latent space.

distribution of the supine breaths that is much more similar to the left, right, and prone position distribution.

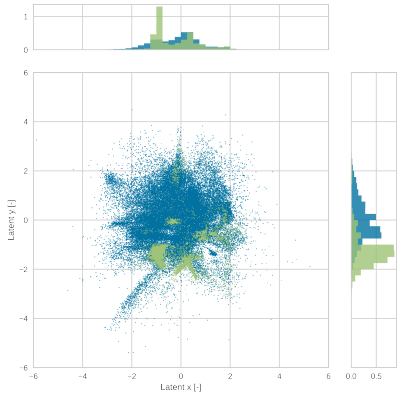


Figure 5.9: Paradoxical breathing events (green) vs. the population distribution (blue) in the latent space.

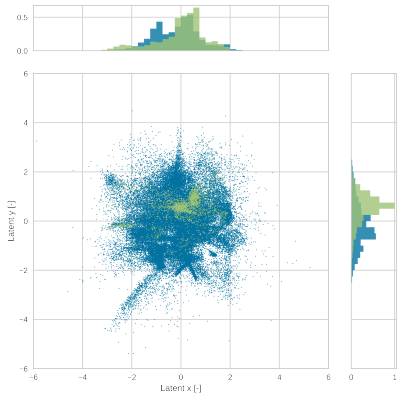


Figure 5.10: Breaths drawn in the prone position (green) vs. the population distribution (blue) in the latent space.

5.6 Discussion

In the previous sections, we both described and presented the results of training a VAE to reconstruct respiratory signals from individual respiratory cycles, resulting in a model able to represent individual breaths with only two values.

Existing work that targets similar focal points and focuses on applying ma-

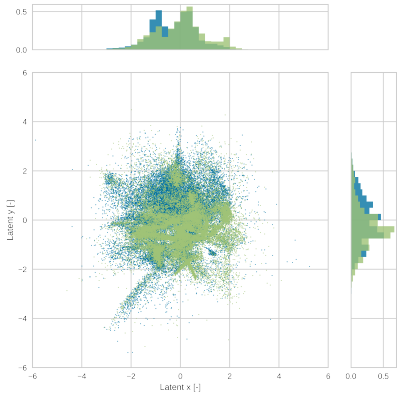


Figure 5.11: Breaths drawn in the supine position (green) vs. the population distribution (blue) in the latent space.

chine learning on respiratory data is mainly rooted in classification of respiratory events, [52]–[54]. However, other applications exist [62]. Our contribution differs from theirs in several ways. In order to shed light on the differences, we will go through the three research questions one by one.

Regarding the first research question, our data indicates that respiratory events, in particular, group together in the latent space, which indicates that some events have a tendency to cluster together, while other events, such as breaths drawn in the supine position (see Fig. 5.11) show no discernible difference in their distribution across the latent space. This illustrates that event types scored by sleep experts seem to cluster together to some extent in the latent space.

Regarding the second research question, our results show that the VAE can accurately reconstruct not only the signals but also the interplay between them. Furthermore, when the RIP belts are in paradoxical movements, the nasal airflow signal shows a clear limitation as well. This indicates that despite the breaths in the latent space being 384 times smaller than the input data, the VAE was able to learn the interplay between the signals. This result supports the intuition that providing the context of the signals (in this case individual breaths for respiratory signals) can allow machine learning models such as VAE to learn the mechanics of signal data significantly easier compared to when using arbitrary segmentation of the data. Additionally, breaths drawn in the prone position showed a drastically different distribution than the other positions, suggesting that the prone position has a much greater effect on breathing than other positions.

Regarding the third research question, we show that despite the fact that the

reconstruction loss during training for the test data was approximately 15% per sample in the signal, the reconstructions visually look exceptionally convincing. A partial reason for a relatively high error rate may be attributed to the model removing high-frequency signal components from the data. We conclude that the VAE can accurately reconstruct the respiratory signals.

The disagreement between scorers for respiratory events is also hinted at by our results as can be seen in Fig. 5.5. Despite the hypopneas that were detected in the RIP signal being second in terms of cluster association, the hypopneas detected in the flow signal were extremely spread across the clusters. Since the hypopneas detected in the flow signal outnumber the hypopneas detected in the RIP signal 6.67 to one, the relatively low agreement makes sense. The low agreement in the scoring for OSA is also visible in our results.

Due to the lack of existing comparable work in this field, and since most comparable works introduce some form of classification into the training process [58], [62], it is difficult to compare this work to prior literature without extending the scope of this paper. Due to the high uncertainty in the scoring for respiratory events in sleep medicine we consider the contributions made in this paper, an advantage over the use of semi-supervised approaches for this type of application.

A limitation of this work relates to the KL regularisation term in the loss function of the VAE, the latent space has significant overlaps. Some works remedy this by simultaneously training a classifier on the latent space, making for a semi-unsupervised approach. This additional classifier training, however, introduces some bias to the data, and as we wanted to keep our model completely unsupervised, we opted to not train the classifier. As mentioned earlier, the reconstructions ignore the higher-frequency components of the signals, which may theoretically contain information on pathophysiology. In its current form, the duration of the respiratory cycle is ignored. Future attempts may find it appropriate to include the length of the breaths in the model.

5.7 Conclusion

This paper describes the results of applying a variational autoencoder machine learning model on respiratory signals from individual breaths, an application that is to our knowledge novel. We performed clustering analysis using KMeans clustering to analyse the tendency of breaths affected by events as labelled by sleep experts, to cluster together in the latent space. We saw that some events, in particular respiratory events, show a tendency to cluster together, indicating that the neural network successfully learned some manifestation of the events in the signals. We also show that even though the latent representations are 384 times smaller than the input data, the reconstructions produced by the decoder VAE components are accurate, producing reconstructions that despite lacking the higher-frequency

components of the input signals which theoretically could be pathophysiologically significant, still simulate the interplay between the signals.

Future work derived from seeds of this work could include further analysis into the latent space, as well as experimentation with larger models. Furthermore, it would be of interest to investigate if sequential breaths during sleep show a similar position in the latent space, for example if a sequence of breaths drift towards the nasal airflow artefact cluster before the signal shows the artefact. Additionally, it was noticed during the work that when the VAE was used to encode and then decode breaths that displayed artefacts, the reconstruction showed that the VAE had attempted to reconstruct the most probable shape of the unusable signal. This reconstructive property suggests that future work could investigate whether VAE or other AE models can be used to recover some information from signals otherwise unusable due to artefacts.

Chapter 6

An optimized framework for processing multicentric polysomnographic data incorporating expert human oversight

6.1 Introduction

The emergence of explainable artificial intelligence (XAI) presents vast potential for revolutionizing various application areas, such as in healthcare [121]. However, despite the great potential, there are significant issues that need to be tackled before XAI can be fully utilized [122]. One such issue originates from application areas within healthcare, where automation of manual tasks and data-driven decision-support has to take the central stage before XAI can become a viable option [123].

A subfield of healthcare is the collection and analysis of sleep recordings, referred to as polysomnography (PSG) [108]. A PSG is an overnight recording of various biomedical signals, such as an electroencephalogram (EEG), electromyogram (EMG), electrooculogram (EOG), and various respiratory signals. Upon collection, the PSG must be manually annotated by a sleep technologist which is a cumbersome and time-consuming task [108]. PSG scoring is a vital step in the process of identifying and diagnosing the presence of many sleep disorders, some of which are extremely prevalent [5]. A sleep technologist will manually review the recording according to a set of rules devised by the American Academy of Sleep Medicine (AASM), labeling events such as respiratory events, and sleep stages in

a process referred to as scoring. The sleep stage scoring is done by assigning a specific class to each 30-second segment (also called epochs) in the recording. The sleep stages classes are categorized into 5 categorical values: the Wake (W) class for wake period, the rapid eye movement category (REM) and three non-REM stages (N1, N2, and N3) that respectively describe the depth of sleep. A product of the PSG scoring is the creation of a hypnogram, a graphical representation tracing the progression of sleep stages throughout the night. This visual tool, often complemented by a hypnodensity graph, provides a detailed overview of the patient's sleep architecture, capturing transitions between sleep stages [124], [125]. Self-applied-PSG (henceforth referred to as simply PSG), a newly designed simplified version of traditional PSG, utilizing frontal EEG instead of the conventional International 10-20 System, refers to a type of sleep study that the participant can set up themselves and sleep with at home for up to three nights in the current work [126].

One of the main drawbacks of the current scoring process is, as stated earlier that it can be excessively time-consuming, which can cause considerable delays in providing sleep reports to healthcare providers and consequently delay diagnosis [127], as well as increase the cost of healthcare considerably [128]. Adding to this challenge, significant inter-scorer variability exists [129]; disagreements can reach 19.3% for sleep stages [129] and 11.6% for respiratory events [130]. Delays and disagreements such as these can have negative effects on patient outcomes, as untreated sleep disorders can have a significantly negative impact on patient health [131].

The advent of machine learning and other automatic scoring algorithms offers a potential solution by automating the process of manual scoring, which the AASM sees great potential in [132]. However, the development and application of machine learning are often prohibitively technical, requiring diverse knowledge of computer science to achieve [133], [134]. There is also a dire need for socio-technical alignment, i.e. the multi-disciplinary collaboration between the computer scientists integrating the algorithms, and the professionals working in the context in which the algorithms are being integrated [134]. The integration of AI, machine learning, or advanced data-driven decision-making of any kind into the workflow may move the industry professionals from a generative role (creating the outputs themselves) to the role of auditors, where they correct the output of the algorithms, and consult with computer scientists to tweak and alter the models to handle edge cases or incorrect generations by the algorithm [135]. Moreover, in the rare case when socio-technical alignment is reached, trust issues often surface, where the professionals working within the context that the algorithms are integrated into, may not trust the outcomes [136], which has posed a great limitation in healthcare [122], [137]. This mistrust has received limited focus in terms of research contributions and needs to be studied further.

Machine learning models are often deemed a 'black box,' owing to their lack of

transparency and the extensive technical knowledge needed to understand them. Moreover, their incapacity to adapt to dynamically evolving requisites often leads to their obsolescence. This has resulted in the increasing prevalence of human-in-the-loop AI systems [138]. Human-in-the-loop AI systems allow one or more human experts to take an active part in the training process by continuously evaluating the model and providing new inputs that are then selectively used to re-train the model in a process called active learning [139].

To advance and modernize sleep research as well as to enable the collection of a large-scale European sleep recording dataset, the Sleep Revolution project, a joint venture involving 24 European partners, was initiated [126]. Each partner contributes approximately 60 sets of three-night PSGs. Sleep technologists then evaluate these on a shared workstation which is a part of the Sleep Revolution high-performance cluster. After this, healthcare professionals analyze sleep parameters, which helps them to diagnose the patient. A significant objective of the Sleep Revolution is to reduce scoring time [126]. One strategy to achieve that goal is to direct the focus of the sleep technologists to the areas of sleep that automatic algorithms have less 'certainty' of. By displaying these areas of high uncertainty, referred to as *gray areas* from now on, we can specifically target the sleep technologists towards these areas, instead of unilaterally trusting or mistrusting the automatic scoring algorithms [95].

Most of the research done on automatic sleep staging algorithms mainly focuses on the increase in model prediction accuracy or agreement. With recent datasets mobilizing an ensemble of independent sleep technologists scoring the same record, research on uncertainty quantification, such as gray area identification in the domain of sleep staging, is growing [96], [129], [140]. However, the union of sleep staging algorithms, including selectively focusing the attention of sleep technologists using uncertainty or gray areas during sleep scoring, is a newborn concept that needs to be assessed.s

To enable these algorithms to benefit sleep technologists in their daily work, a system is required that bridges the gap between the data collection and the manual scoring itself. To collect the data required for this work, a digital platform was designed to handle automatically collecting, segmenting, and processing the PSG. The concept of digital platforms takes into account that a digital platform is both a piece of software, while it is also an intermediary that connects needs with resources. Therefore the concept of digital platforms encompasses a larger array than the software itself as it, in a socio-technical manner, also takes the context into account. In this case, the digital platform is accessed via the users' web browser and is hereinafter referred to as the platform.

Computer-assisted automatic scoring with manual review has demonstrated the ability to reduce PSG scoring time significantly, with some studies showing improvements by factors of 1.26 to 2.41[71]. Moreover, automatic sleep scoring algorithms can halve the scoring time [72], [73].

Some research on the integration of automated scoring has been conducted in the last few years, as listed in Table 6.1. Rayan et al. (2023) [24] discuss the challenges and advancements in automatic sleep scoring in the context of rodent and human sleep research. They note limitations in handling atypical data and lack of flexibility but also note that automatic algorithms can make the process more efficient. A recent study evaluated a deep-learning-based automatic scoring software for its accuracy and efficiency compared to manual scoring. The results indicated a high correlation between the automatic scoring system and manual scoring, particularly in sleep staging and the apnea-hypopnea index. The automatic scoring system also demonstrated a significant reduction in manual scoring time, leading to improved workflow efficiency in sleep laboratories [73]. Oxholm et al. (2021) [74] interviewed 9 healthcare professionals and 5 patients about their attitudes towards using data from electronic health records in an algorithm to screen for alcohol abuse in hospitals. Professionals were mixed in their views, appreciating the tool's time-saving potential but concerned about losing instinctual decision-making. While this work is only tangentially related to our work, the authors point out the requirement to include healthcare professionals in the process of integrating automatic algorithms. Gerla et al (2018) [75] presented a computer-assisted approach for sleep staging using EEG recordings and AASM 2012 scoring rules, focusing on real clinical data with artifacts and missing electrodes, evaluating the influence of AI in clinical settings by comparing traditional manual sleep stage classification with AI-based methods, including expert-in-the-loop strategies, for the analysis of EEG recordings in sleep studies. In a later study, Gerla et al. (2019) [76] developed a semi-supervised method for evaluating PSG, blending expert-scored segments with automated classification. This approach, tested on both healthy individuals and chronic insomnia patients, showed enhanced efficiency and accuracy in sleep data analysis compared to conventional manual scoring methods, demonstrating the impactful role of AI in streamlining sleep study workflows.

6.1.1 Contributions

As is evident from Table 6.1, existing research on automatic sleep scoring addresses either the impact on workflow or the opinions of medical professionals on AI in the workflow. To the best of our knowledge, no research exists that addresses the integration of automatic sleep scoring into existing work environments which is an important aspect to consider to achieve socio-technical alignment.

To fill this research gap, we designed both a platform and a process for evaluating the effectiveness of introducing gray areas into the work of sleep technologists and their trust in the process. By integrating the platform featuring machine learning algorithms into the work of sleep technologists through our empirical case within the Sleep Revolution, we extrapolate three main contributions. Firstly, we outline the architecture for a platform that has been designed and developed to

Table 6.1: Comparison of contributions of this work and similar work.

Work	**Addresses integration into existing work environments**	**Addresses impact on workflow**	**Addresses opinions of medical professionals on AI in workflow**
Our work	X	X	X
[24]		X	
[73]		X	
[74]			X
[75]		X	
[76]		X	

enable the integration of automatic scoring. Secondly, we introduce the concept of ”gray areas” as a method of selectively focusing the attention of sleep technologists on fewer areas in the PSG. Thirdly, we illustrate the decreased scoring time and increased agreement gained by integrating the automatic scoring algorithms into the workflow of sleep technologists. Throughout this research, and particularly when analyzing the results, we realized that the phenomena we encountered consistently and that was common to all of our results, was missing a clear clinical terminology that we attempt to address in this work.

6.2 Materials and methods

PSG sharing and scoring between research centers require sophisticated architectures that rest heavily on the principles of storing and processing medical data cohesively. The proposed platform has the main purpose of connecting needs with resources, which in this case outlines the sharing and scoring of PSG between research centers.

The methodology is three-fold; 1) the design and development of the platform, 2) the validation of the platform, and 3) interviews with sleep technologists. The design section covers the architecture, components, and technologies chosen to implement the platform, the validation section covers how the platform was assessed in terms of processing duration, sleep technologist speed, and agreement improvements, and the interview section describes how sleep technologists were interviewed for their sentiment toward integrating AI tools into the workflow.

6.2.1 Platform design

The platform needed to be conceived in agreement with the main constraints as having a simple user interface for sleep technologists to be able to authenticate and upload their PSG; providing administrative oversight on uploads from different centers; being fault-tolerant; and being scalable [141]. The platform is split into three distinct components:

1. Web-based front-end for user uploads, administration, and dispatching of jobs to the other components (henceforth referred to as the *front end*).
2. Three-night PSG splitter (henceforth referred to as the *splitter*).
3. Processing pipeline that augments PSGs with automatic scorings (henceforth referred to as the *processor*).

Figure 6.1 shows an overview of the platform architecture. An important feature of the platform is to allow users (e.g. sleep technologists and healthcare professionals) to upload multiple PSGs to be shared and scored at the same time without breaking the platform. To achieve this, the FastAPI Python web framework was used, which despite its simplicity handles multi-user web applications supporting asynchronous code [142]. The platform is protected with a user login access in which each user is a validated member of Sleep Revolution consortium[143].

Additionally, the front end handles receiving signals from both the splitter and the processor via HTTP requests and issuing jobs to the splitter when a new PSG is received and to the processor when a PSG has successfully been split. Splitting is necessary when several nights' PSG are combined into one file. The job queue was achieved using a RabbitMQ queuing server, which is a program that allows disparate asynchronous programs to communicate by listening and issuing messages to a queue [144]. By utilizing a message-queue protocol, the font end can offload more time and memory-consuming projects such as generating automatic scorings to other processes, thus reducing the probability of users experiencing downtime, or data loss.

The processor is the final component of the architecture. Its purpose is to prepare the individual night PSG by augmenting the PSG with the AI scoring, along with the gray area scoring. The output of the processor is twofold. Firstly, the processor prepares a 'scoring' version of the PSG that is augmented with predicted sleep stages from an automatic scoring algorithm integrating gray areas and is made available for manual scoring, and a version meant for later computer processing and machine learning. Each component was containerized using the virtualization software Docker [145] for enhanced isolation, consistency, and reproducibility during deployments, which is important in sustainable and secure development.

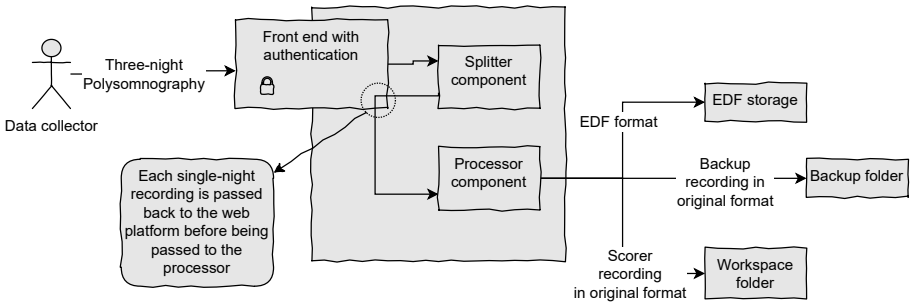
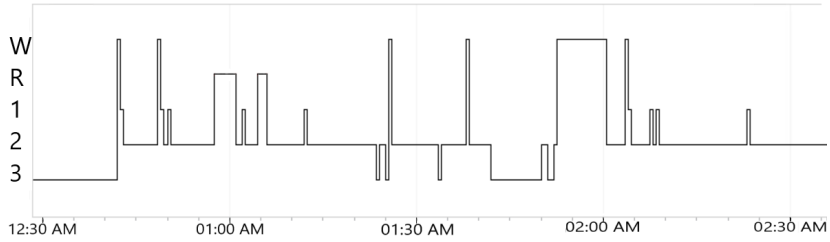


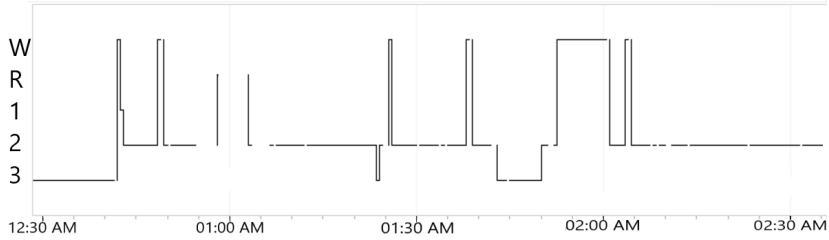
Figure 6.1: Overview of the platform showing how the front end, processor, and splitter are combined.

As introduced previously, the processor prepares the PSG to be manually scored, stored, and ready for further analysis. To reduce the manual work of the sleep technologists, a crucial step in the processor is highlighting areas in the PSG that are hard to score for the algorithm, i.e. gray areas. The gray area augmentation works first by sending each one-night PSG EDF file to the trained deep learning model aSAGA [96]. The aSAGA architecture is based on a revisited U-time architecture for scoring and respiratory events prediction [91], [146]. The U-time is an encoder-decoder structure consisting of blocks of consecutive convolutional, batch normalization, and pooling layers. However, in the aSAGA algorithm, a single-channel model is used, which was first trained on PSGs' EEG (C4-M1) and then fine-tuned with an EOG (E1-M2) channel using self-applied PSGs with frontal setup. This was done to have generalizability between EEG and EOG channels and to increase the compliance of the model for frontal EEG and EOG setups. The aSAGA model is parameterized to return a hypnogram of the same length as the number of epochs from the signal input. The model has an accuracy of 80% estimated over different scored sleep datasets. This accuracy is on par with manual scoring [129], however, the gray areas from aSAGA model prediction have been validated by comparing the match with the gray areas from predicted manual scoring uncertainty.

The second part concerns the gray areas. Using the predicted hypnodensity from the aSAGA model as input, a trained clustering algorithm tags each epoch that belongs to the gray areas [95]. The clustering algorithm is a multi-objective method based on multinomial mixture models clustering the different levels of sleep technologist agreement and summarizing the results into two sets of high-agreement and *gray area* clusters. The threshold is selected according to the maximization of the distance between two distributions of the sleep technologist's agreement measure. When the algorithm receives a new hypnodensity, it outputs a hypnogram called aSAGA-UA with *gray area*. Figure 6.3b illustrate such predicted hypnogram where each *gray area* are represented as a line discontinuity



(a) aSAGA predicted 2 hours hypnogram. W: Wake; R: Rapid-Eye-Movement; 1, 2 and 3 are respectively for N1, N2 and N3.



(b) aSAGA-UA predicted 2 hours hypnogram with gray areas. Each discontinuity in the hypnogram line represents a gray area. W: Wake; R: Rapid-Eye-Movement; 1, 2 and 3 are respectively for N1, N2 and N3.

Figure 6.3: Example of output 2 hours hypnograms for the n°1 PSG from 50 × 10 PSG, obtained using the processor and rendered in Nox Medical's Noxturnal software for manual review.

named "whitespace".

Using aSAGA-UA, it becomes easier for the sleep technologists to view epochs where the AI scoring may not be accurate, and need to be re-evaluated. In Figure 6.3a between 12:30 a.m. and 1:30 there are many transitions between Wake, N1, N2, and REM scored by the algorithm. For the same time period in Figure 6.3b, there are many line discontinuities characterized by a whitespace symbolizing gray areas. For instance, the predicted N1 and REM sleep stages in Figure 6.3a are not present in Figure 6.3b where whitespaces are clearly visible instead. Regarding the high number of sleep transitions happening in a few minutes, the associated signal might be hard to interpret by the algorithm. A manual review from the sleep technologist is needed in that part of the hypnogram. The method has been evaluated on a real case of uncertainty analysis of 50 PSGs manually scored by 10 sleep technologists. We refer to this dataset as 50×10PSG. This dataset comes from a cohort of 50 participants that have previously been scored by ten independent sleep technologists to create a consensus scoring [95], [96]. After testing the

clustering algorithm on predicted hypnodensities from aSAGA, the threshold separating the gray area clusters from other epochs was lowered to 0.73 according to a sleep technologist's recommendations. The new value avoids the creation of an excessive amount of white spaces in the final hypnogram.

All three components were hosted on a Linux virtual machine run on a Cisco Hyperflex high-performance compute system located at Reykjavik University. The virtual machine was equipped with 10 Intel(R) Xeon(R) Gold 6248R central processing units, and 20 gigabytes of random access memory.

6.2.2 Sleep technologist time and consensus validation

The platform is validated with the help of three sleep technologists, referred to from this point as Sleep Technologist One, Sleep Technologist Two, and Sleep Technologist Three (ST1, ST2 and ST3 respectively). ST1 and ST2 are experienced sleep technologists, whereas ST3 is considered less experienced. Each sleep technologist was asked to score a randomly selected subset of from the 50×10PSG. The sleep technologists looked at multiple channels when scoring the data, using the NoxTurnal sleep study annotation program. The EOG channels E2 and E3 referenced against AFz were used to track eye movements, the EEG channels AF3, AF4, AF7, AF8 all referenced against the average value of the eye channels E3 and E4 were used to observe EEG activity. There were no EMG electrodes as part of the setup, however the sleep technologists used the muscle component derived from the eye electrodes to their aid in scoring. Each sleep technologist received half of the subset scored with a default proprietary industry-standard automatic scoring and the other half had the automatic scoring with gray areas (aSAGA-UA). We also refer to these options as without and with AI, respectively. The partitioning of the subsets can be seen in Table 6.2.

Table 6.2: The layout of PSGs to be scored, where X indicates default automatic scoring and O indicates aSAGA-UA, that is aSAGA with gray areas. The numbers correspond to specific recordings in the 50×10 PSG.

PSGs	1	2	3	4	5	6	7	8	9	10
ST1	X	O	X	O	X	O	X	O	X	O
ST2	O	X	O	X	O	X	O	X	O	X
ST3	O	X	X	X	X	X	O	O	O	O

The sleep technologists were instructed to score sleep stages and arousals. The PSGs with the default automatic scoring were manually reviewed as sleep technologists would normally do in a clinical setting, reviewing every epoch manually. For

the PSGs with aSAGA-UA, only the gray areas were manually reviewed by the sleep technologists. The standard operating proceeding follows these specific steps:

1. Start by running automatic analysis.
2. Adjust the time frame from lights out to lights on (start and stop times for the correct analysis period).
3. Score sleep stages and arousals according to AASM version v. 3.0 [7].
4. For the aSAGA-UA scoring, after reviewing all visible gray areas, look for possible missed epochs by searching for sleep stage scorings that contain the word "uncertain", and correct them.

Each sleep technologist was asked to accurately measure the duration of the scoring process for each PSG in their subset. Subsequently, their scoring was collected and compared to the existing consensus scoring from the 50×10 PSG. The scoring accuracy of the sleep technologists using the system as support was assessed via Fleiss's multi-rater [147] κ coefficient. This coefficient $\kappa \in [0, 1]$ measures the agreement of the current sleep technologist sequence to the scoring sequences given by the ten sleep technologists in the consensus scoring. In the case of samples with high agreement between sleep technologists, Fleiss's κ coefficient converges to 1 and 0 otherwise.

6.2.3 Interviews with sleep technologists

The perceived trust and reliability of the automated scoring system were evaluated through semi-structured interviews with the three sleep technologists, following an interview guide. These 30-minute interviews aimed to explore the sleep technologists' confidence in the system's output and their comfort in integrating the system into their workflow. The sleep technologists provided feedback on the system's overall performance, as well as reflected on their trust in the system's automatic scoring algorithm and gray area identification. The interviews were transcribed verbatim and relevant segments of the interviews were and the qualitative data was analyzed with thematic analysis.

6.3 Results

This section is divided into three main subsections. Firstly, we present the platform's performance. Secondly, we present the performance gain in terms of both scoring time and agreement of the sleep technologists. Thirdly, we present the results from interviews with sleep technologists.

6.3.1 Platform performance

Table 6.3: Samples of processing time in minutes (min) taken by each queue according to file size in mego octets (Mo).

Splitter		Processor	
File Size (Mo)	Processing time (min)	File Size (Mo)	Processing time (min)
1920	1.8 ± 0.3	640	3.4 ± 0.3
2160	2.1 ± 0.0	720	4.6 ± 0.3
2400	2.0 ± 0.0	800	4.8 ± 1.2

Table 6.3 lists the time taken by the two main components of the pipeline, the splitter and the processor. Since the platform must split the upload into individual nights before further processing, the initial three-night PSG takes approximately 458.5 seconds (7.6 min) to become available for scoring, including both splitting and processing time. However, for the subsequent PSGs, the sleep technologists mainly perceived the processing time, which averages 336.2 seconds (5.6 min) per PSG. The processing time for later PSGs is negligible, as the sleep technologist can begin scoring the first file while the others are being processed. Consequently, the processing time is optimally utilized, preventing any significant delays in the scoring workflow. When employed in the early stages of the data collection, the queue sizes of the splitter and processor did not grow to excessive lengths, with the processor queue generally not exceeding the size of three pending processing jobs.

The front end was designed to be clean and provide a structured interface to submit both the PSG itself and information about the participant. A screen grab of the user interface is presented in Figure 6.4.

6.3.2 Sleep technologist time and consensus validation

This section provides insights into sleep technologists' performance when using the AI for scoring. Each sleep technologist received an identical set of PSGs to score, both with and without aSAGA-UA. Here, the analysis encompasses the sleep technologist's time efficiency and agreement metrics.

In the following section, the study results first delve into the outcomes of the sleep technologist's time to score. Figure 6.4 displays the scoring duration for each of the 10 PSGs and all sleep technologists with and without aSAGA-UA assistance. As seen in Fig. 6.4a, ST1 using aSAGA-UA assistance shows an average ± standard deviation scoring duration of 20.8±8 min compared to 36.8±16 min for ST2. ST1

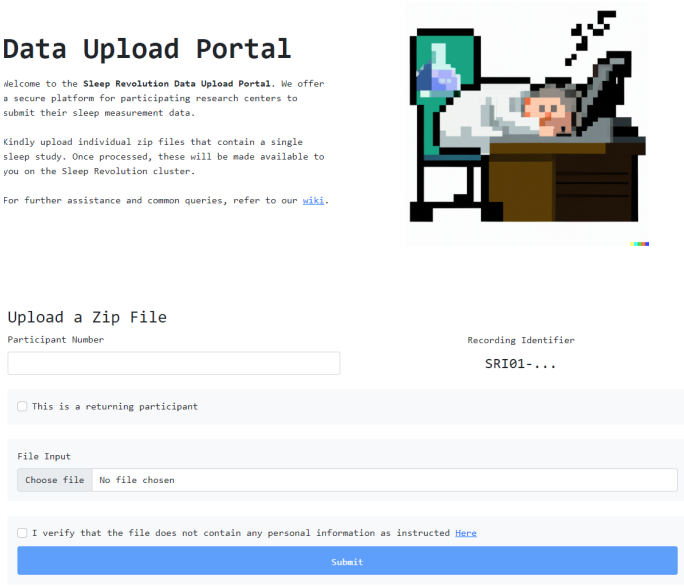


Figure 6.4: Front end user interface for uploading recordings.

reviews faster than ST2. This efficiency translates into an average scoring duration reduction of 16 minutes.

Meanwhile, as seen in Figure 6.4b, when using aSAGA-UA, ST2 approximately equaled the time of ST1. ST2 displayed a scoring duration of 26±9 min, and ST1 displayed 30±6 min, with ST2 reducing their mean scoring duration by 4 minutes when using aSAGA-UA. Finally, Figures 6.4a and 6.4b display ST3 having a time of 111±26.7 min without AI. ST3, as the least experienced in this study, was noticeably slower in scoring than the other sleep technologists. However, Figure 6.4c shows that ST3 depicted a significant decrease in the time to score, of 46±20.2 when using aSAGA-UA, or a reduction of 65 minutes.

Turning to the agreement analysis, Table 6.4 is divided into two parts; the first half details the sleep technologist's agreement based on the analysis of the complete PSGs, while the second half assesses the agreement specifically for the gray area epochs which the sleep technologist handled with aSAGA-UA assistance.

When the agreement was calculated based on gray areas, the sleep technologist using aSAGA-UA assistance was the only one aware of the nature of these epochs. A steady trend in the overall agreement of sleep technologists using AI for scoring is observed in Table 6.4. However, the agreement rating of ST3 appears to be negatively affected by the use of aSAGA-UA assistance. This reduction is possibly attributed to a more challenging sample of associated PSGs, which generally

Table 6.4: Fleiss's multi-rater κ mean±standard deviation estimated on overall hypnograms and gray areas epochs only by sleep technologists manually scoring and using aSAGA-UA assistance.

Sleep Technologist	Complete hypnogram		Gray areas only	
	Without AI	With AI	Without AI	With AI
ST1	0.87±0.05	0.86±0.05	0.76±0.12	0.73±0.08
ST2	0.72±0.08	0.85±0.04	0.48±0.22	0.65±0.17
ST3	0.84±0.04	0.80±0.08	0.60±0.11	0.77±0.12

achieved a lower agreement score, but this is not clear.

Figure 6.5 shows the agreement analysis of each of the three sleep technologists in this study with the 50×10 PSG with and without aSAGA-UA. The performance levels of ST3, represented in Figures 6.5a and 6.5c, consistently matched or surpassed the levels achieved by ST1 and ST2 while using the aSAGA-UA tool. However, there is a discernible decrease in scoring agreement for the three PSGs illustrated in Figure 6.5e when ST3 utilizes aSAGA-UA. Despite this decrease, the performance levels of ST3 generally remained comparable to and occasionally exceeded those of ST2. Moreover, across all three figures (6.5a, 6.5c and 6.5e), the agreement of the sleep technologists stayed consistent, indicating that the scorings produced traditionally and using aSAGA-UA are comparable. For instance, PSGs IDs 1, 7, 8, and 10 all display complete agreement, having been scored twice by the sleep technologists using aSAGA-UA.

A second aspect depicted in Table 6.4 is the agreement of the sleep technologists only calculated for the gray areas on both samples (sleep technologists using aSAGA-UA assistance and without aSAGA-UA assistance). In this table, ST2 and ST3 got a marked increase in their agreement when using aSAGA-UA, with ST2 and ST3 gaining approximately 0.17 κ, but with ST1 a decrease of 0.03 κ for the gray areas.

Figures 6.5b, 6.5d and 6.5f offer the Fleiss's multiraters κ estimated on only gray areas per PSG ID. In this comparison, only the sleep technologist using aSAGA-UA knew that these epochs were labeled as gray areas. As expected, in Figures 6.5b, 6.5d, and 6.5f, ST1 depicted the same stability observed previously. ST2 and ST3 showed an increase in κ when using aSAGA-UA. The increase, however, was less strong for ST2 who showed more agreement's dispersion among the PSGs. Otherwise, in Figure 6.5f, ST3's agreement showed a significant increase for PSGs 1 and 9 compared to ST2's agreement which might explain the difference in Figure 6.5e. This last result indicates that aSAGA-UA assistance may benefit a beginner

more than an experienced sleep technologist.

In summary, all the participating sleep technologists showed a decrease in their time to score but to a different degree. Regarding their scoring agreement, the sleep technologists depicted three distinct results when using aSAGA-UA. The agreement of the experienced sleep technologist with the PSG signals was not affected by aSAGA-UA. On the other hand, the second experienced sleep technologist has shown more dispersion among the PSG with on average an increase when using aSAGA-UA. Finally, the third, less experienced sleep technologist benefited the most from aSAGA-UA assistance.

6.3.3 Interviews with sleep technologists

The sleep technologists were interviewed about their experience using aSAGA-UA, and the interview transcripts were compiled into a word cloud (Figure 6.6).

Initially, the sleep technologists approached the new system with optimism. ST1 expressed initial enthusiasm: "[Before starting] I was very optimistic that it would decrease the scoring time". All sleep technologists found it simple to integrate AI scoring with gray areas into their current workflows with ST3 commenting "I do not think it is an issue at all [...] it is pretty easy to implement".

However, as they used the new system, the sleep technologists noticed a need for a more accurate staging algorithm, with ST1 noting "What I saw is that the algorithm is not good enough". For ST1 and ST3, improved accuracy is essential for reducing the scoring time and building trust in the new system. ST2 provided a slightly different perspective, suggesting that the system's staging accuracy might already be on par with the inter-scorer agreement of human sleep technologists.

Overall, all three sleep technologists expressed in various ways that trust in AI technology is significant for its continual adaptation into their practice. The sleep technologists articulated the psychological impact of integrating AI staging with gray areas into their workflows. ST2 expressed concern that the AI suggestion might slightly shift their bias in selecting a sleep stage. ST3 spoke along similar lines: "Maybe I had an unconscious bias to lean towards the [suggested sleep stage]".

Overall, the sleep technologists found the new system promising and were optimistic about the approach. ST3 was interested in seeing a more detailed quantification of the gray area uncertainty, asking for "the percentage of the prediction or something like that". However, all sleep technologists agreed that improving the staging algorithm's accuracy was important, as ST1 put it: "You need to trust the algorithm". Their sentiments reflected a cautious optimism, recognizing the potential benefits while anticipating enhancements in usability and trust as the accuracy of the underlying staging algorithm improves.

6.4 Discussion

6.4.1 Main contributions

This paper introduces an advanced web platform aimed at filling the gap of sharing, processing, and storing three subsequent nights of PSG in the sleep research field. The platform has three distinct components: a front end, a PSG splitter, and a processor component with automatic scoring and storing of each PSG. The front end is connected with the two subsequent parts using a flexible message-queue protocol, preventing the front end from crashing in case of failure in the processing of PSGs. The platform was tested on a set of 60 three-night PSGs files. The average processing time of the platform ranged between 5.6 min, for an associated file size of 1920 Mo, and 7.6 min, for a file of size 2400 Mo.

Moreover, the automatic scoring, including the gray areas implemented in the processor component has been assessed with the help of three sleep technologists. The predicted scores by the platform showed a decidedly positive effect on the speed of scoring. This enhancement is achieved without significantly complicating the workflow of sleep technologists. The strategic incorporation of AI support into their routine not only optimizes the time efficiency of scoring but also adds a layer of precision and reliability to the process. The most experienced sleep technologists showed a high agreement on an average of 0.85 κ when using AI support. This value of agreement is in line with the observed agreement obtained for other data sets manually scored [148]. Additionally, a significant increase in both the scoring speed and agreement was observed for the less experienced sleep technologist, suggesting that the use of automatic algorithms and gray area assistance has the potential to bridge the gap between more experienced sleep technologists and the less experienced ones, and thus speeding up the training of new sleep technologists.

6.4.2 Platform insights

Utilizing a message queue protocol imparted a considerable complication in implementing the platform that would have been avoidable if we had instead opted for a separate process using e.g. HTTP requests, or implemented the splitting and processing as part of the same program as the front end. Utilizing message queues in favor of more ad-hoc solutions allowed us more flexibility and scalability than with other solutions. The need to split PSGs similarly complicated the work, since it added a component to the process. However, the benefits gained from working with separate nights later in the process outweighed this added complexity.

In the results part, the processor component has been evaluated over a study composed of three sleep technologists with different experiences scoring 10 PSGs with and without aSAGA-UA. However, the dispersion obtained in the results re-

flected a lack of PSG required to obtain an accurate representation of the time to score and sleep technologists' agreement distributions. A study with a greater number of PSGs would allow us to validate the result obtained in the presented paper. Moreover, using aSAGA-UA, the effectiveness of the sleep technologists in terms of scoring duration is affected differently. Their disparity may be explained by the difference in experience with the self-applied PSG frontal signals, the baseline speed of both sleep technologists, and the trust given to the AI-predicted scores in the gray areas. Furthermore, a study with a higher number of PSGs and more sleep technologists is needed to have a better estimation of the effectiveness obtained by the use of AI as a scoring support tool.

The interviews revealed the sleep technologists' agreement that the platform integrated well into their workflow, with ST3 commenting especially on the ease of implementation. The sleep technologists did raise issues with the performance of the scoring algorithm itself, with ST1 reporting that the scoring algorithm is "not good enough". ST3 expressed some concern that the sleep stage recommendation system was influencing their decision-making. This worry reflects the need for trust and alignment between the sleep technologist and the algorithms, especially in the context of healthcare AI recommendation systems. As the final sentiment of ST3 indicates, the experts display interest in having more insight into the reason why the algorithm assigned areas as gray, aligning with the rise in demand for xAI, reflecting a broader desire for transparency and clarity in human-in-the-loop AI systems.

6.4.3 Clinical Acquiescence of AI

Traditional accuracy and agreement measures are both derived from the confusion matrix offering an overview of the performance of the classification algorithm. Accuracy variation across different datasets of less than 1% is considered insignificant for that kind of algorithm [96], [149]. However, confusion matrix-derived metrics such as accuracy only assess if the algorithm prediction matches the correct output. It does not guarantee that the algorithm captures a key signal pattern related to a specific sleep stage hiding in this 1% accuracy variation. For clinical experts, such as sleep technologists, it is crucial to ensure that key signal patterns are correctly interpreted. If a scoring algorithm with high accuracy and agreement is missing these key patterns, it becomes hard for the sleep technologist to trust the algorithm's prediction. To summarize, there is a need for a metric assessing the scoring algorithm's conformity that also assures sleep technologists' trust in the algorithm. Clinical acumen is a term symbolizing the ability of healthcare professionals to make quick and accurate decisions on complex issues that a clinical AI along with a human-in-the-loop might include in the future to make a diagnosis [150]. In this work, we would like to introduce a general term to define the act of

accepting or agreeing to the use of AI as a decision-making tool by clinical experts: Clinical Acquiescence.

6.4.4 Study limitations

Our research is not without limitations and below we highlight the most notable ones. Although the web platform was architectured with the main purpose of being scalable and robust, this paper does not include an extensive scalability evaluation of the web platform itself. In the current study, this was not the focus, as the platform was tested and evaluated primarily on the improvement it could provide in the task of scoring PSG. Future works could be directed toward stress-testing the platform, evaluating the maximum number of PSG it can handle simultaneously, and determining whether the web platform could sustain heavy traffic loads without considerably slowing down or crashing. Since neither the splitter nor processor queue grew to prohibitively big lengths during testing, we did not see a reason to implement scaling functions, nevertheless, the implementation of the system as a whole lends itself well to dynamic scaling. The gray area threshold of 0.73 was selected with the help of a sleep technologist, and adjusted to produce the least number of gray areas without including blatantly incorrect algorithmic scorings. This study does not evaluate the effect of this threshold on scorer speed or reliability and notes that the threshold value is highly dependent on the algorithm used to produce the scoring. Detailed sensitivity analysis would need to be performed on the threshold value in order to evaluate its performance and create guidelines on how to optimally determine its value. Only global metrics such as the scoring duration and the agreement of the sleep technologists have been considered in this paper. However, this study does not go into detail about the source of the uncertainty in sleep staging between sleep technologists. For instance, it is well known that one primary uncertainty source is the transition between the sleep stages N2 and N3 [95], [140].

6.4.5 Future work

In the future, a replication of this study needs to be performed, with a greater number of both sleep technologists and a larger subset of PSG to gain a broader perspective of the effects of integrating AI augmentation into the sleep technologist's workflow, along with algorithm trust assessment.

The next step would be to loop the manual review of the gray area with the automatic scoring algorithm. This process is referred to as active learning [139], [151], and aligns with the AI-integrated human-in-the-loop workflow. A continuous loop would link the reviewed gray area with the scoring AI updating the model and sending a new set of gray areas corresponding to the actual sleep technologist.

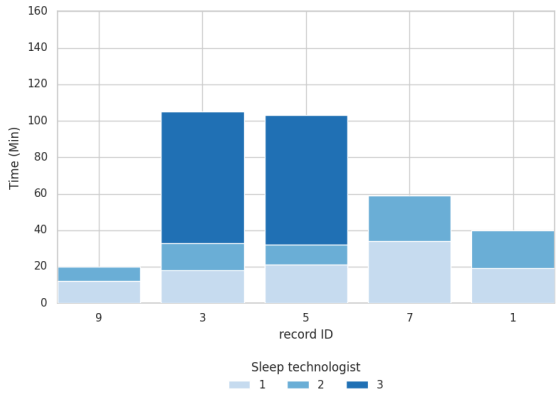
Due to the modularity of the platform, it is easy to add more algorithms and augments to the processor, making the adoption of any additional algorithms more approachable without resulting in downtime or causing data loss. For example, the BreathFinder [152] respiratory isolation algorithm is planned for addition to the processor to allow future analysis of individual respiratory cycles. Additionally, adding new destinations and output formats for the PSGs is made easy, e.g. using a micro-scoring platform with integrated machine-learning capabilities, currently under development.

One possible avenue to further advance the platform is to allow researchers to upload their custom automatic scoring algorithms to be vetted and be run autonomously on test data, without ever having to gain physical or digital access to the data, allowing for a reliable method for testing disparate algorithms on the same datasets for greater consistency, reproducibility and transparency in future sleep research.

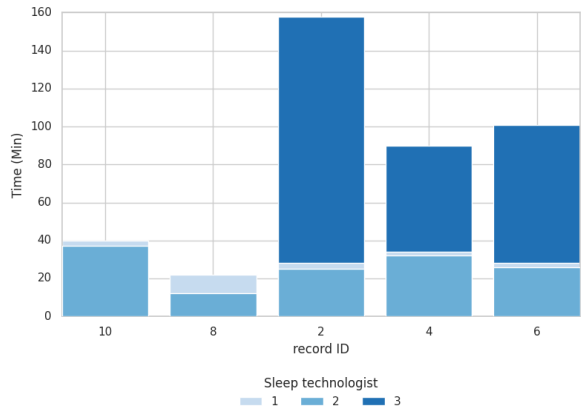
6.5 Conclusion

In this work, we presented a platform that enables PSG collection, integrated with automatic AI scoring algorithms. We evaluated the platform in terms of its effect on sleep technologists' time, and accuracy when scoring PSGs that incorporate AI assistance. In our restuls, we observed a clear gap in research addressing the integration and evaluation of automatic scoring algorithms for PSG.

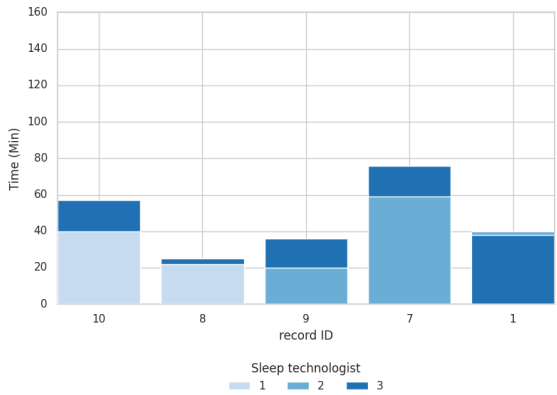
The proposed platform incorporates AI assistance but still prioritizes the human expert as the ultimate decision-maker. This balance of human expertise and AI presents a promising avenue for future advancements in the field of sleep study and analysis, potentially leading to more refined and accurate diagnostic practices.



(a) Scoring duration with **ST1** using pipeline aSAGA-UA assistance.

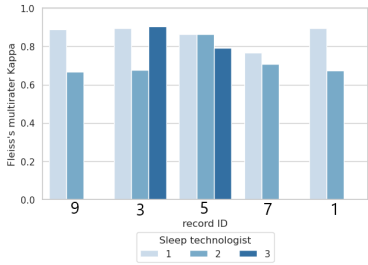


(b) Scoring duration with **ST2** using pipeline aSAGA-UA assistance.

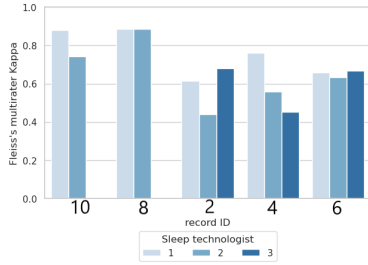


(c) Scoring duration with **ST3** using pipeline aSAGA-UA assistance.

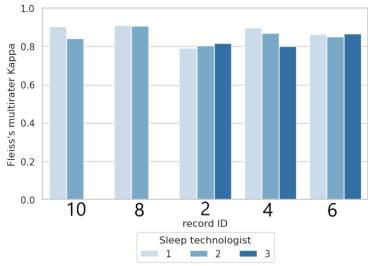
Figure 6.4: Overlapped bars of scoring duration comparison of PSG with one sleep technologist using aSAGA-UA and the other two using the standard procedure.



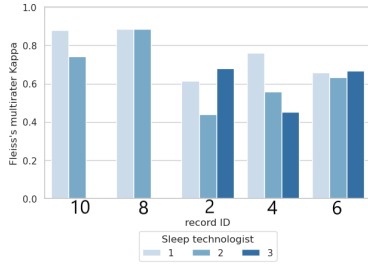
(a) Fleiss's multiraters κ overview with **ST1** using aSAGA-UA assistance.



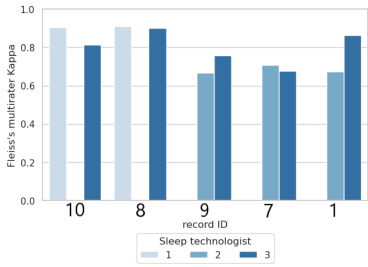
(b) Fleiss's multiraters κ of only gray area epochs with **ST1** using aSAGA-UA assistance.



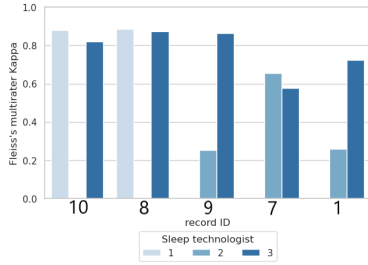
(c) Fleiss's multiraters κ overview with **ST2** using aSAGA-UA assistance.



(d) Fleiss's multiraters κ of only gray area epochs with **ST2** using aSAGA-UA assistance.



(e) Fleiss's multiraters κ overview with **ST3** using aSAGA-UA assistance.



(f) Fleiss's multiraters κ of only gray area epochs with **ST3** using aSAGA-UA assistance.

Figure 6.5: Agreement analysis of 10 sleep technologists compared to a sleep technologist with or without aSAGA-UA assistance. The first column **(a,c,e)** shows the total agreement per polysomnographs. The second column **(b,d,f)** shows the agreement for gray area epochs tagged by the artificial intelligence.

Figure 6.6: A word cloud of the interview transcripts.

Chapter 7

World of ScoreCraft: Online Multi-Scorer Consensus Study on DSS in Sleep Staging

7.1 Introduction

As sleep disorders and sleep issues are extremely prevalent in society [4], [153], [154], sleep technologists are more important today than ever. The sleep technologist is responsible for reviewing sleep studies, or polysomnograms (PSG), and manually annotating (also known as scoring) sleep stages and events such as movement, arousal, and apnea [8]. Medical doctors subsequently use this analysis for diagnosis.

A PSG is an overnight collection of biometric signals, such as various respiratory signals, electrooculogram (EOG), electroencephalogram (EEG), and electromyogram (EMG), to name a few [8]. The scoring of a PSG is a time-consuming task, which takes up to two hours to score a single eight-hour PSG [24], and can significantly strain the sleep technologist and induce scoring fatigue. Alongside being laborious and time-consuming, the scoring of a PSG can also vary significantly between sleep technologists, with disagreement on sleep staging being as high as 14% [129] and respiratory events as high as 34.6% [11], and obstructive apnea severity classifications differing in 66% of cases depending on the scorer [155].

To reduce the need for a sleep technologist to set up and monitor a traditional PSG and to eliminate the requirement to sleep in a sleep laboratory overnight, new PSG equipment has been developed for home use. This multichannel frontal PSG enables patients to apply the device themselves and sleep comfortably in their own beds. These self-applied PSG devices are currently undergoing validation and a recent device demonstrated a success rate of approximately 90% in enabling effective

self-application and accurate data collection [156].

The rise of artificial intelligence (AI) has not left the scoring process behind, with various automatic algorithms being designed to automate and speed up the work of sleep technologists by detecting sleep stages [157] and apnea [158], [159] to name a few tasks. These algorithms have great potential to assist in the scoring process; however, they cannot be treated as a drop-in replacement for the human expert since they are incapable of adapting and adjusting to the evolving standards of care as the human experts are. They also need to be rigorously trained for different issues that may arise in diverse application contexts and be aware of possible biases that occur in AI based on sex, age, comorbidities, and other factors [160].

To meet the need for systems that aim to utilize AI as a tool for the human expert instead of as a replacement, decision support systems (DSS) have been designed to provide interactive tool sets to assist experts in making decisions and solving unstructured or semi-structured tasks Sprague [161]. Garg, Adhikari, McDonald, *et al.* [162] found that in 64% of cases, DSS or similar systems considerably impacted clinician performance. Articles on DSS have been widely written in the field of sleep research; however, the literature covers mostly automation of individual tasks, but does not examine the effect of the inclusion of DSS into the workflows of sleep technologists in terms of accuracy or time taken to score sleep recordings. Furthermore, the effects of integrating DSS and AI into sleep technologists' workflows can provide significant advantages in terms of speed and accuracy, but to be integrated effectively requires building trust towards the AI [163].

Along with the aforementioned effects, an important aspect to measure is the potential impact of the DSS in changing the behavior of the human expert or how the professionals whose toolset is augmented with AI might, for example, become complacent and default to the AI recommendations [164]. Complacency towards AI can be defined as a tendency of the user to not appropriately scrutinize the results of the automated tools. The tendency towards complacency is complicated to analyze but has been shown to be linked to the transparency of the AI system, as well as how well the expert expects the AI to perform [165]. Another important aspect of integrating AI is the concept of 'clinical acquiescence,' defined by Holm, Jouan, Hardarson, *et al.* [166], which refers to the willingness to adopt AI assistance in clinical workflows.

There is a noticeable gap in the literature on the effects of integrating DSS into the workflows of sleep technologists. Most of the existing literature focuses on the accuracy of the algorithms designed to automate sleep-scoring tasks [97], [166]. However, the impact assessment of such algorithms on expert performances is a key component that is usually missing. We propose to leverage this by studying the changes introduced by human experts whose toolsets have been augmented with DSS. In more detail, we aim to investigate the effects of integrating AI into the work environment.

This work investigates the effects of introducing recommendations in scoring

sleep stages. We further measure the effects of the recommendation presentation and correctness on the accuracy and speed of the human expert.

This study used a repeated measures design with two conditions to collect a consensus scoring for one hour of traditional and self-applied PSG. The main conditions being researched were the effect of recommendations presence and study type, as the objective of this study is to research the effects of recommendations on the sleep staging process. The study also examines the effect of the type of scoring recommendations (human or AI) on the scoring process, counterbalancing the recommendations by only showing recommendations for one session of each PSG type. Hence, 3 factors were being studied: type of sleep study, presence of recommendations, and type of recommendations. Their significance is measured in terms of scoring accuracy or correctness and time.

7.2 Methodology

For this study, the scoring sessions were limited to a single hour (120 epochs), chosen from a data set that was recorded simultaneously using traditional PSG and self-applied PSG equipment. The hour was chosen for good-quality signals, and the hypnogram featured multiple transitions between sleep stages. The sessions were limited to one hour to focus on collecting a greater diversity of scorings and to prevent scoring fatigue from affecting sleep technologists during the process.

7.2.1 Platform

The scoring collection was performed with the MicroNyx online scoring platform [167], which allows secure online scoring of PSGs. The MicroNyx platform enables measuring difficult-to-obtain features, such as the decision-making time, change of mind, and more aspects of the scoring process. In preparation for this study, multiple sleep technologists were recruited before the experiment to validate and provide recommendations and feedback on the scoring interface and signal filtering to measure the impact of the new MicroNyx system on the scoring in a Co-Design process. This was repeated several times over a few months until each sleep technologist at Reykjavik University could reliably score with 80% consensus with a pre-existing scoring, or roughly around the 86% expected agreement of sleep technologists [129]. MicroNyx allows for the creation and the scoring of so-called 'scoring sessions,' variable-length signal segments containing signals that sleep technologists can then score for research purposes. The signals, filtering, and data source can be customized and tailored to different research purposes.

7.2.2 Recommendations

In 50% of epochs, participants were presented with a recommendation for a sleep stage. The recommendation rate of 50% was chosen to provide an equal amount of epochs with and without scoring recommendations. To investigate the presence of any potential bias against automatic algorithms, the recommendations were sourced from a human scoring for the PSG and self-applied PSG studies, respectively. For consistency, the same sleep technologist performed the scoring for the PSG and self-applied PSG recommendations (blinded to which recordings were from the same participants). Despite all scoring being sourced from the same human sleep technologist, the recommendations were either presented as being from a human sleep technologist or an automatic AI scoring system. This is henceforth referred to as the recommendation presentation. The ratio of human vs. AI recommendations was equal to ensure equal representation.

Both accurate and deliberately incorrect recommendations were implemented to evaluate the influence of recommendations on scoring behavior. The rate of incorrect recommendations was set to 20%, aligning with the inter-scorer variability reported by Nikkonen, Somaskandhan, Korkalainen, *et al.* [129] and emulating the error rate of sleep technologists. To ensure realistic incorrect recommendations, a sleep-stage map was developed using results from a multicentric consensus scoring study [129], which identified the most common misclassifications by sleep technologists. Figure 7.1 illustrates the most common misclassifications, providing insight into how incorrect recommendations were designed to reflect typical scoring disagreements.

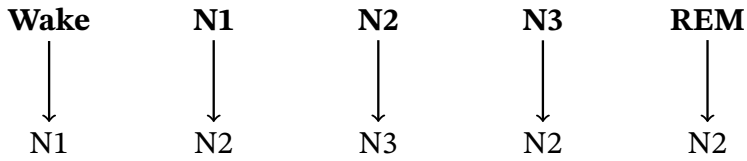
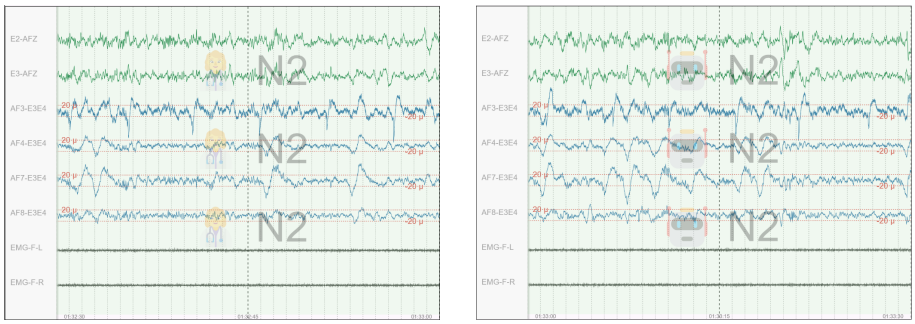


Figure 7.1: Map of sleep stages to deliberate misclassification for scoring recommendations.

Recommendations presented as being from humans were represented with a scientist emoji (see Figure 7.2a). In contrast, the recommendations we presented as being from an AI had the robot emoji, as shown in Figure 7.2b. The recommendations were designed to be noticeable for sleep technologists without being intrusive or obscuring the signals meaningfully. The post-study survey included questions about the visibility of the recommendations to ensure that the participating sleep technologists could easily spot the recommendations and tell the difference between human and AI recommendations.



(a) Recommendation presented to be from a human

(b) Recommendation presented to be from an AI

Figure 7.2: Comparison of recommendations presented as human vs artificial intelligence (AI)

7.2.3 Data setup

The data for this study was selected from a 'double-setup' dataset, where a traditional PSG recording along with a self-applied PSG setup was placed on the participants, and the two types of PSG were recorded simultaneously [97].

The sleep technologists scored both types of PSG using the MicroNyx web scoring interface, which supports flexible and customizable signal selection and filtering. Scoring guidelines for required signals and their appropriate filtering were followed to ensure consistency with established methodologies and software for both traditional and self-applied PSG. This subsection is divided into two parts, detailing the signals and filtering options for traditional PSG and self-applied PSG, respectively.

Traditional PSG

The traditional PSG setup presented to the sleep technologists was based directly on the AASM recommendations for scoring sleep stages, utilizing the appropriate signals and filters as listed in the guidelines [8]. An illustration of the PSG signals is provided below.

The signals the sleep technologists used to score the traditional PSG were the EEG signals C4-M1, C3-M2, F4-M1, F3-M2, O1-M2, O2-M1, the EOG signals, E1-M2, E2-M1, and the chin EMG.

The EEG signals were filtered using a 0.5–35 Hz bandpass filter. Each EOG signal was processed through a 0.3 -35 Hz bandpass filter and sampled at 200 Hz. The chin EMG signal was passed through a 10 Hz high-pass filter and sampled at 200 Hz.

Figure 7.3 shows how the signals from the traditional PSG were presented to the sleep technologists in the MicroNyx scoring interface.

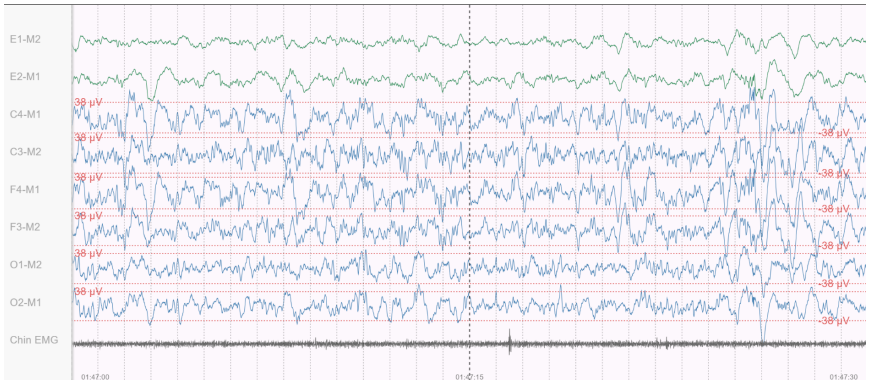


Figure 7.3: The MicroNyx scoring interface displaying a traditional PSG

Self-applied PSG

The self-applied PSG signal used by the sleep technologists to score the self-applied PSG were the EEG signals AF3-E3E4, AF4-E3E4, AF7-E3E4, AF8-E3E4 and the EOG signals E2-AFz, E3-AFz. The EEG signals were filtered using a 0.5-35 Hz bandpass filter and sampled at 200 Hz. The EOG signal was processed through a 0.3-35 Hz bandpass filter and sampled at 200 Hz.

Since the self-applied PSG setup did not include a traditional chin EMG, the E1 signal referenced against the E3 signal, along with the E2 signal referenced against the E4 signal, were used as stand-ins for the EMG signal, with a 10 Hz high-pass filter applied to produce left and right EMG signals, respectively.

Figure 7.4 shows how the signals from the self-applied PSG were presented to the sleep technologists in the MicroNyx scoring interface.

7.2.4 Reference standard

After the scoring collection, a so-called reference standard was created, a majority-vote scoring for each epoch, which could be used to compare scorings. To have a stable, more reliable hypnogram to compare scorings to than the single-scorer hypnogram, the user scorings for the traditional PSG sessions that did not have recommendations present were used to create a consensus scoring referred to henceforth in this work as the reference standard. This was done to alleviate two major issues. The first issue is that the single-scorer hypnogram was created on different software and is thus not perfectly comparable to the scorings generated on the

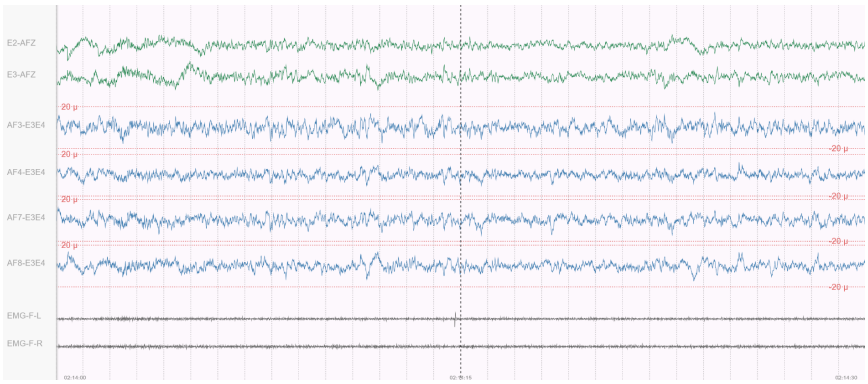


Figure 7.4: Self-applied polysomnography in the MicroNyx scoring interface

platform used for this study. The second issue that the reference standard solves is that the scoring of a self-applied PSG is not as validated as a traditional PSG. By creating a consensus scoring for the traditional PSG that can be compared against the scorings for the self-applied PSG, the difference in the scoring of traditional and self-applied PSG is more pronounced. This work uses the reference standard to compare the participant scorings and the recommendations. We refer to sleep technologists' scorings as *accurate* and recommendations as *correct* if they respectively match the reference standard. The words are not considered interchangeable in this work since recommendations cannot be considered 'accurate,' as they are not themselves scorings.

7.2.5 Recruitment of participants

The study was heavily dependent on the participation of sleep technologists. To recruit sleep technologists as participants in the scoring collection, an email invitation was sent to 49 sleep technologists affiliated with the Sleep Revolution [126] project. In addition, the study was advertised at conferences in the European Respiratory Society (ERS) and the European Sleep Research Society (ESRS) in September 2024. An invitation was also emailed to European Society of Sleep Technologists members via a newsletter. In total, 16 sleep technologists across Europe participated in the study.

7.2.6 Study procedure

Each technologist was instructed to complete a 10-minute (20-epoch) tutorial session in which they were shown how to navigate the scoring interface, how recommendations appeared, and how to score using the MicroNyx interface successfully.

After the tutorial, each technologist was directed to complete two scoring sessions, one for PSG and one for self-applied PSG, where they scored exactly one hour of sleep data per session. After those two sessions, the sleep technologists were instructed to wait a week before scoring two additional sessions. The waiting time was instructed in order to prevent familiarity with the recordings.

After successfully completing the four sessions, a link to a short post-study survey was presented to the sleep technologists. The post-study survey aimed to gauge how each sleep technologist perceived the MicroNyx platform for its ease of use and ability to score sleep stages on the scoring interface. The post-study survey questions are provided in appendix .1.

The MicroNyx platform ensured that each participant completed one traditional PSG session and one self-applied PSG session in randomized order, followed by a one-week break, after which they repeated the process with a second pair of sessions. Recommendations were presented in only one of the two traditional sessions and one of the two self-applied sessions, ensuring that if recommendations were provided in the first traditional session, they would not be shown in the second traditional session, and the same rule applied to the self-applied sessions. Recommendations are covered in more detail in the following section.

7.2.7 Analysis

Data analysis was performed using Python and R, with tools chosen to suit the specific requirements of each test. In Python, the Pandas library [168] was used for data preparation and organization. For simpler single-variable analyses, the SciPy library [169] was employed to conduct hypothesis testing through paired T-tests, assessing significant differences in decision-making time and accuracies under different conditions, assuming normality. When the normality assumption was unmet, the Mann-Whitney test [170] was applied as a non-parametric alternative. A significance level of $alpha = 0.05$ was used throughout. To investigate the relationship between scoring variables and decision-making time, the R programming language [171] and the ARTool library [172] were utilized to perform an Aligned-Rank-Transform Analysis of Variance (ART ANOVA). ART ANOVA was chosen for its ability to accommodate the continuous nature of the decision-making time alongside categorical predictors such as recommendation correctness, presentation style, and PSG type, which do not satisfy the assumptions of standard ANOVA. A generalized linear model was applied to analyze scoring accuracy, accounting for the binomial distribution of the dependent variable. This approach enabled the evaluation of categorical predictors' main effects and interactions, such as recommendation correctness, presentation style, and PSG type, while respecting the constraints of binary data.

7.3 Results

This section presents the results of this study, including the aggregate performance of sleep technologists when scoring traditional and self-applied PSG and the granular effects of recommendations on scoring accuracy and time. The analysis highlights differences between user-level and epoch-level outcomes, focusing on the impact of recommendation correctness and presentation. Cumulatively, the 16 participating sleep technologists completed 64 scoring sessions, producing 9158 individual scorings for the total 240 epochs in the scoring sessions. To create a reference standard, a majority vote approach was taken using the scorings of the 16 participants. This was used to assign a unique label to each epoch. After using a majority-vote system to create the reference standard, the scorings from the original hypnogram used to source the recommendations aligned with the reference standard with 76.23% accuracy; however, due to the intentional 20% error rate, the agreement of the recommended sleep stages with the gold standard dropped to 67.57% for the traditional PSG and 57.64% for the self-applied PSG.

When analyzing the decision-making time, the time per epoch was obtained by calculating the time difference of the creation timestamps of successive scorings. A log transformation was applied to the time-per-epoch data to remove statistical outliers, and an interquartile range filtering with a threshold of 1.5 was used to remove unrealistically long decision-making times. The filtering step identified 598 outliers, with a mean time-per-epoch of 3210.6 seconds and a standard deviation of 77646.7 seconds, indicating that the filtered values deviated significantly from the remaining decision-making times. The remaining data used for the analysis had a mean scoring-per-epoch time of 2.0 seconds and a standard deviation of 1.9 seconds.

7.3.1 Participants

In total, 16 sleep technologists participated in the study, successfully completing four scoring sessions. Of the 16 sleep technologists, 13 answered the post-participation questionnaire. The questionnaire (see appendix .1) showed that most sleep technologists felt confident in their ability to interpret and score signals, with over 84% selecting 6 or 7 on the confidence scale. Two participants (7.7% each) selected 2 or 5, indicating some variation in confidence levels. Of those who answered the questionnaire, all reported that they could easily see the recommendations and that it was easy to see if they were from a human or AI. The sleep technologists were from 11 different countries: Australia, Belgium, Finland, France, Germany, Guatemala, Iceland, Ireland, Italy, Portugal, and Spain.

Of the various recruitment methods, the invitation sent to the relevant members of the Sleep Revolution [126] yielded six sleep technologists, and the letter

sent to the ESST yielded five. Two sleep technologists participated after hearing about the study from colleagues or during a talk where the study was advertised.

7.3.2 Aggregate analysis of scoring accuracy and time

To establish a baseline standard for scoring accuracy and time without the influence of recommendations, we filtered the data to exclude sessions that included recommendations from the analysis. When comparing the scoring accuracy of sleep technologists, 11 out of 16 technologists performed better in scoring the traditional PSG sessions than in scoring the self-applied PSG sessions (see in Figure 7.5 the accuracy achieved by each sleep technologist in both PSGs). Conversely, five sleep technologists scored the self-applied PSG sessions more accurately than the traditional PSG sessions. Overall, sleep technologists demonstrated a tendency for a slightly higher accuracy when scoring the traditional PSG sessions, achieving an accuracy rate of 85.7% compared to 81.0% for the self-applied PSG. The difference in accuracy was, however, not statistically significant (paired t-test: p=0.098, rank-sum: p=0.15).

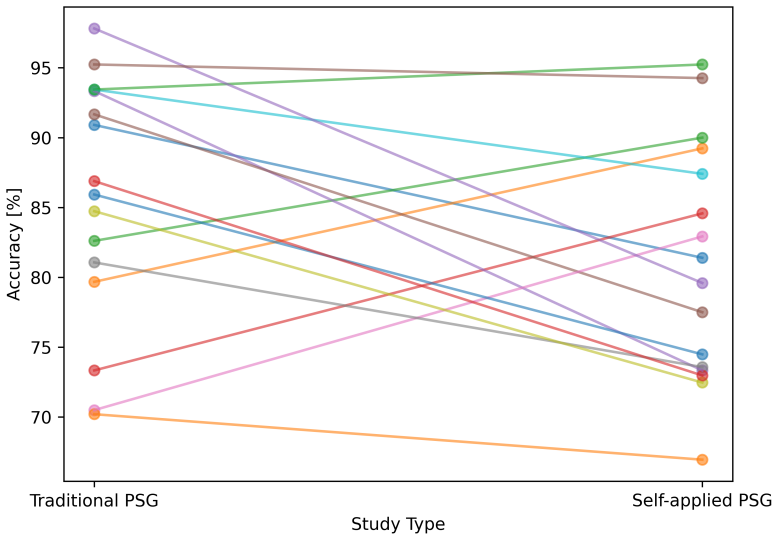


Figure 7.5: Sleep technologist change in accuracy between traditional and self-applied PSG. Each line represents one sleep technologist.

The decision-making time was analyzed similarly to the scoring accuracy. For

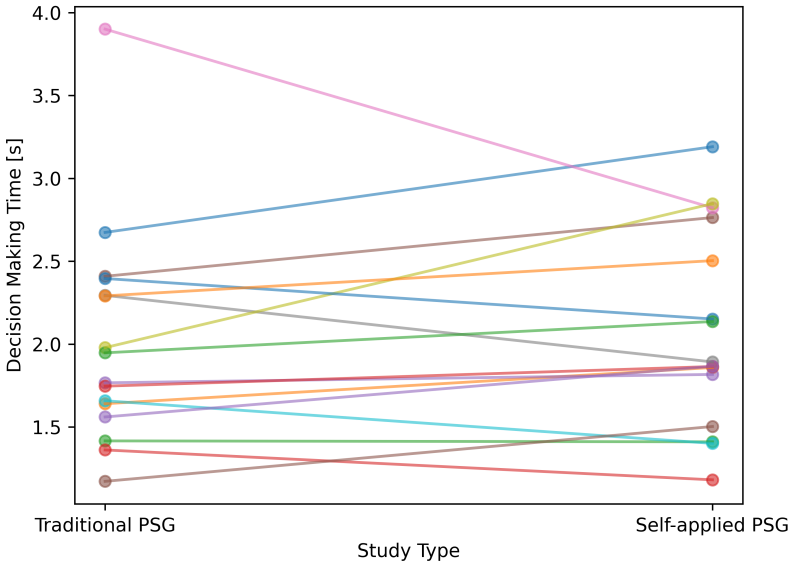


Figure 7.6: Sleep technologist change in decision-making time between traditional and self-applied PSG. Each line represents one sleep technologist.

the traditional PSG, the average decision-making time was found to be 2.0 seconds, and for the self-applied PSG, 2.0 seconds and is displayed in Figure 7.6, not statistically significant (paired t-test: p=0.58, rank-sum: p=0.74).

7.3.3 Epoch-Level effects of recommendations on accuracy

The sessions for both traditional and self-applied PSG were separated based on whether or not they included recommendations, and the scorings were then compared in terms of accuracy with the reference standard. The effect of correct recommendations on the overall accuracy of scorings for all scoring sessions can be seen in the heat map in Figure 7.7. Notably, the difference between accuracies for the self-applied PSG and the traditional PSG was more dramatic, with the baseline accuracy for self-applied PSG being 81.47%, but decreasing by approximately 3.2% to 78.26% when recommendations were present. The accuracy difference was not statistically significant for the traditional PSG (t-test p=0.73, rank-sum p=0.73). However, for the self-applied PSG, the difference was found to be statistically significant (T-test p=0.006, rank-sum p=0.006).

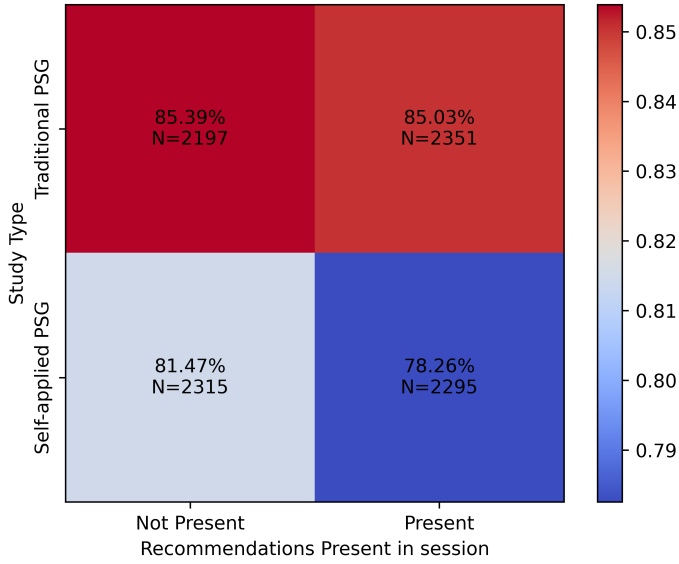


Figure 7.7: Effect of recommendation presence on scoring session accuracy for traditional vs. self-applied PSG.

When studied further by separating scorings based on their recommendation presence and correctness in Figure 7.8, the effect of correctness on recommendation becomes clearer. The baseline overall scoring accuracy on an epoch-by-epoch basis without recommendations remained at 85.38% for the traditional PSG sessions and 81.47% for the self-applied PSG sessions.

When faced with incorrect recommendations, sleep technologists assessed the traditional PSG with an accuracy of 82.13%, which is 3.26 percentage points lower than the baseline accuracy of 85.39% achieved when scoring the traditional PSG without any recommendations. This effect is even more pronounced for the self-applied PSG, with the accuracy rate decreasing from the baseline 81.47% achieved when scoring the self-applied PSG without recommendations to 74.20% when faced with incorrect recommendations, making for a 7.27% decrease in accuracy. Correct recommendations had the opposite effect on scoring accuracy, with sleep technologists achieving 90.76% accuracy when scoring traditional PSG epochs featuring an accurate recommendation, resulting in a 5.37% increase in accuracy from the baseline scoring accuracy without recommendations. This positive effect was also observed for the self-applied PSG, where sleep technologists achieved 88% ac-

curacy when presented with correct recommendations, up 6.53% from the 81.47% baseline scoring accuracy for self-applied PSG without recommendations.

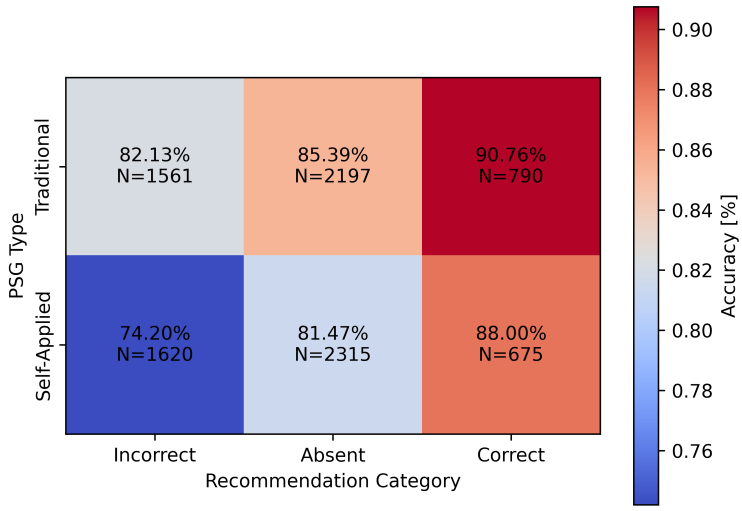


Figure 7.8: Effect of recommendation correctness on scoring accuracy.

The generalized linear model results can be seen in Table 7.1, which displays odds ratio (OR) change for each change in variables from the baseline (intercept) where the study type is a traditional PSG, the presentation is human, and the recommendation is correct. The baseline odds ratio is 8.25, meaning that when study type, presentation, and correctness equals the baseline, the accuracy of sleep technologists is 89.18%. The presentation and study type did not affect the scoring accuracy with statistical significance. However, the recommendation accuracy, on its own, had a significant effect on the scoring accuracy, lowering the accuracy of sleep technologists from the 89.18% baseline to 82.54%. The interaction between presentation and study type did not significantly affect the scoring accuracy, with AI recommendations when scoring self-applied PSG lowering the accuracy from the baseline by 10.32% to 78.85%. The final interaction that statistically significantly affected the scoring accuracy was for self-applied PSG when recommendations were incorrect, which lowered the scoring accuracy to 72.34%, or by 16.83%.

When plotted in a three-way line plot (see Figure 7.9), the interactions from Table 7.1 become more clear. For both traditional and self-applied PSG, incorrect

Table 7.1: Generalized linear model linear regression results. The three-factor interaction term was included at first but was not significant. Thus, it was removed from the model.

	OR	2.5%	97.5%	Significance
Intercept	10.492	7.543	14.595	*
C(Presentation)[T.AI]	1.215	0.776	1.901	
C(StudyType)[T.Self-applied]	0.856	0.555	1.321	
C(Correctness)[T.False]	0.379	0.244	0.588	*
C(Presentation)[T.AI]:C(Study Type)[T.Self-applied]	0.661	0.404	1.082	
C(Presentation)[T.AI]:C(Correctness)[T.False]	1.062	0.654	1.723	
C(StudyType)[T.Self-applied]:C(Correctness)[T.False]	0.561	0.342	0.919	*

recommendations decreased the scoring accuracy in line with the results from Figure 7.8. However, the negative impact of incorrect recommendations was not as dramatic for the traditional PSG as it was for the self-applied PSG. The presentation of correct recommendations had a paradoxical effect on accuracy with respect to the study types. For traditional PSG, human recommendations produced a mean accuracy of 90.43%, and AI recommendations produced a mean accuracy of 93.64%. Meanwhile, for self-applied PSG, human recommendations produced an average accuracy of 90.99%, and AI recommendations produced a mean accuracy of 86.77%.

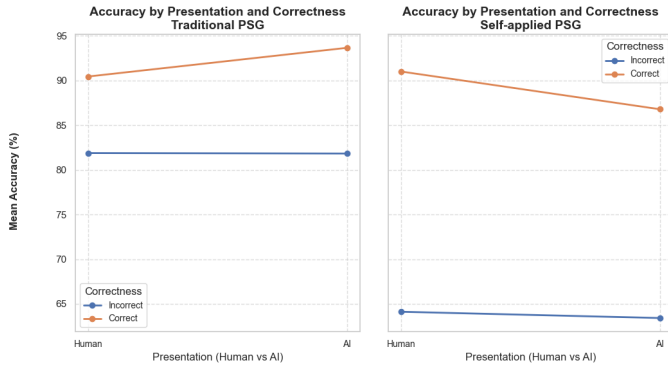


Figure 7.9: Three-way line plot with grouped comparisons between effects of study type, recommendation presentation, and recommendation correctness on scoring accuracy.

7.3.4 Epoch-Level effects of recommendations on decision-making time

The effects of recommendations on decision-making time were analyzed using a method similar to the effects on accuracy. When separated based on study type and recommendation presence (see Figure 7.10), sleep technologists spent 1.9 seconds scoring per epoch on average without recommendations, which rose to 2.0 seconds per epoch when scoring with recommendations. This effect was on the boundary of statistical significance (T-test p=0.041, rank sum p=0.077).

The average time per epoch for self-applied PSG showed a similar trend, with sleep technologists spending 2.0 seconds on average per epoch when scoring without recommendations, which rose to 2.1 seconds when recommendations were introduced. Unlike traditional PSG, this effect was statistically significant (T-test p=0.0036, rank-sum p=0.0039).

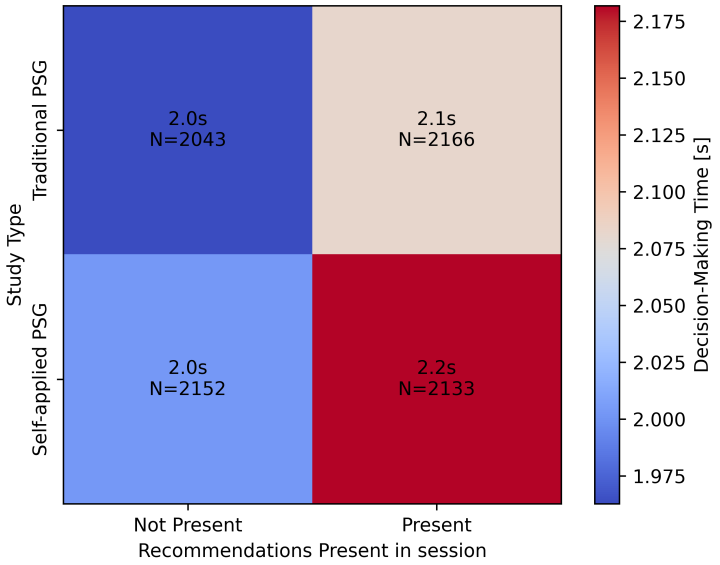


Figure 7.10: Effect of recommendation presence on decision-making time for traditional vs. self-applied PSG.

Similar to the scoring accuracy, this effect was clearer when scorings were separated based on recommendation correctness and presence (see Figure 7.11). For both traditional and self-applied PSG, the sleep technologists spent 2.0 seconds

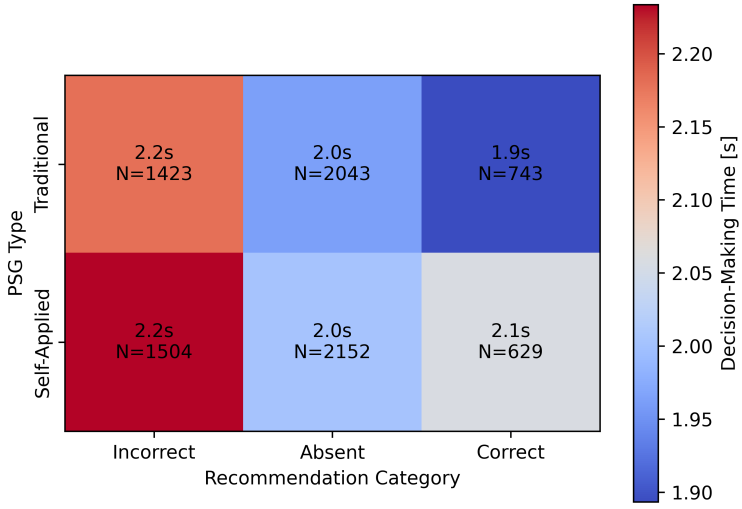


Figure 7.11: Effect of recommendation correctness on decision-making time.

on average per epoch when not faced with any recommendations. For incorrect recommendations, the sleep technologists spent 2.2 seconds per epoch on average for both types of PSG. When recommendations are correct, however, the average time-per-epoch decreased by 0.1 seconds for the traditional PSG; however, the decision-making time for self-applied PSG increased by 0.1 seconds on average.

Similarly to the accuracy, ART ANOVA was used to discover which features of the recommendations and their interactions affected the decision-making time. The results of the ANOVA can be seen in Table 7.2, where the significance levels are marked with one to three asterisks, depending on the significance level. The effect of the presentation alone was insignificant for the decision-making time, and its interaction with the correctness and PSG type did not reach statistical significance. PSG type was found to affect the decision-making time with statistical significance, and its interaction with the correctness of the recommendations had a borderline significant effect on the decision-making time. The correctness was highly statistically significant for the decision-making time. The three-way interaction of the variables had a highly statistically significant effect.

The three-way interaction between study type, recommendation presentation, and recommendation correctness revealed distinct trends in decision-making time

Table 7.2: ART ANOVA results for presentation, study type, and correctness on average decision-making time.

Effect	Df	Df.res	F value	Pr(>F)
Presentation	1	2150	0.257	0.612
PSG Type	1	2150	6.860	0.008 **
Correctness	1	2150	38.702	5.915e-10 ***
Presentation:PSG Type	1	2150	4.361	0.036 *
Presentation:Correctness	1	2150	0.837	0.360
PSG Type:Correctness	1	2150	0.391	0.531
Presentation:PSG Type:Correctness	1	2150	16.193	5.917e-05 ***

(Figure 7.12). For traditional PSG, for epochs featuring correct human recommendations, the sleep technologists spent an average of 1.99 seconds per epoch and 1.85 seconds for epochs featuring AI recommendations, or approximately 0.04 seconds shorter when the recommendations were presented as being from AI. For self-applied PSG, correct recommendations displayed a similar trend of sleep technologists taking less time on average to score epochs with AI recommendations (2.0 seconds) vs. human recommendations (2.3 seconds).

Incorrect recommendations showed considerable increases in average decision-making time per epoch, as stated earlier, with sleep technologists spending on average 2.9 seconds scoring epochs with an incorrect human recommendation vs. 2.2 seconds for an incorrect AI recommendation. Self-applied PSG reverses this trend; the sleep technologists spent 2.5 seconds on incorrect human recommendation epochs vs. 3.0 seconds on incorrect AI recommendations.

7.4 Discussion

The main contributions of this work are threefold:

1. We found no significant difference in the scoring accuracy between traditional and self-applied PSG.

2. We found that correct recommendations increased the scoring accuracy for both the traditional and self-applied PSG up to approximately 90% accuracy.

3. We found no evidence for bias toward AI recommendation over Human recommendations when scoring sleep stages.

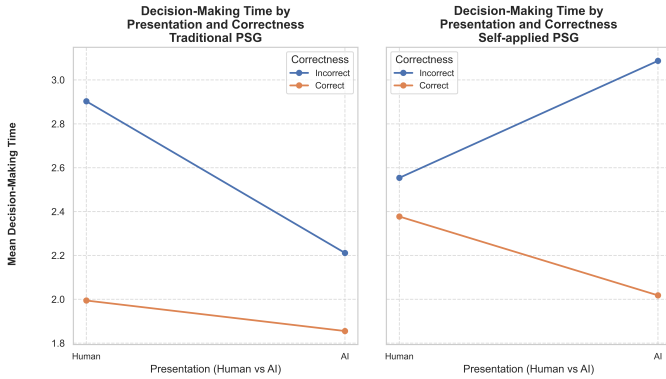


Figure 7.12: Three-way Line plot with grouped comparisons between the effect of study type, recommendation presentation, and recommendation correctness on decision-making time.

The findings of this study provide strong evidence for the potential of AI and DSS to enhance sleep stage scoring by improving accuracy and reducing decision-making time. This work contributes significantly to integrating AI-driven and automated scoring systems into the workflows of sleep technologists, paving the way for faster processes and more precise diagnostics.

Traditional vs. self-applied PSG

When analyzed, no evidence was found that sleep staging accuracy differed for traditional or self-applied PSGs. While the participants were slightly more accurate when scoring traditional PSG epochs, a likely explanation for this is that the majority of sleep technologists participating had not scored a self-applied PSG until in this study. Along with the scoring rules for self-applied PSGs being less defined than for traditional PSGs those two factors are the most likely to affect the scoring accuracy. Our findings align with Rusanen, Korkalainen, Gretarsdottir, *et al.* [97], demonstrating that self-applied PSGs can be reliably scored without additional time or accuracy penalties compared to traditional PSGs. Furthermore, we did not find that the baseline decision-making time differed between the study types, suggesting that the time taken to score self-applied PSGs is not greater than for traditional PSGs.

Scoring accuracy

The results showed that the baseline scoring accuracy without recommendations was in line with the work of Nikkonen, Somaskandhan, Korkalainen, *et al.* [129],

as sleep technologists achieved a comparable baseline agreement with the prior work exploring the inter-scorer variability in sleep staging. Sleep technologists were more likely to score self-applied PSG epochs incorrectly when presented with incorrect recommendations than with traditional PSG epochs. Both types of PSGs showed reduced scoring accuracy with incorrect recommendations; however, self-applied PSGs suffered a significantly greater decrease relative to its baseline accuracy, with a 7.26 percentage point drop compared to a 3.26 percentage point drop for traditional PSGs. This discrepancy is potentially due to sleep technologists being generally less familiar with self-applied PSG signals, thus deferring the scoring decision to the recommendations. While Rusanen, Korkalainen, Gretarsdottir, *et al.* [97] found that self-applied PSGs tends to suffer from noisier signals than the traditional PSGs or electrode placement issues, this does not apply to this study since the scoring session periods were chosen for their signal quality.

Whether recommendations were presented as being from humans or AI had seemingly no statistically significant effect on the accuracy, indicating a high clinical acquiescence. This finding underscores the potential for seamless integration of AI tools in clinical workflows.

These results have significant clinical implementations, as both the traditional and self-applied PSG scoring show significant improvements in terms of accuracy when correct recommendations are integrated into the scoring process.

Decision-making time

As Table 7.2 shows, when epochs with recommendations are analyzed, the study type becomes a significant factor for decision time, with sleep technologists spending an average of 0.2 seconds less time scoring traditional PSG epochs. The correctness was also found to be significant for the decision-making time, with sleep technologists spending 0.3 seconds longer scoring epochs with incorrect recommendations for the traditional PSGs and 0.1 seconds longer for incorrect self-applied PSG recommendations.

The interaction between study type and recommendation presentation also influenced the decision-making time, with sleep technologists spending considerably more time evaluating incorrect AI recommendations for self-applied PSGs than similar recommendations for traditional PSGs. This disparity suggests that, due to their familiarity with traditional PSGs, sleep technologists could more swiftly dismiss incorrect AI recommendations while dedicating additional effort to understanding incorrect human recommendations. Conversely, sleep technologists scored incorrect human recommendations for self-applied PSGs faster than AI recommendations. This indicates that while technologists were more inclined to defer to human recommendations, they subjected AI recommendations to greater scrutiny, reflecting a nuanced approach to clinical acquiescence in the context of novel applications.

The clinical implications for these findings are significant as the reduction in time observed in our results was 0.3 seconds for the traditional PSG epoch, or approximately 13% of the mean decision-making time, meaning that if an eight-hour PSG is scored in two hours, with recommendations, the scoring time will become approximately 103 minutes or a 17 minute gain in time. However, this gain in scoring time is not observed to such a degree for the self-applied PSG epochs, with the time reduction being 0.1 seconds on average, resulting in a 4% speed up, resulting in a scoring time of 114 minutes or a gain of five minutes.

Study limitations

Despite the promising contributions, the study has several notable limitations. First, the scoring sessions were limited to one hour, or 120 epochs, to reduce participant burden. While practical, this time restriction resulted in a relatively small representation of sleep stages, which may limit the applicability of the findings to full-night studies. Additionally, the one-hour duration was not designed to observe fatigue-related effects that might influence scoring accuracy over extended periods, of interest for future studies, as well as whether scoring is affected by the time of day it is being performed. Second, although the participant pool of 16 sleep technologists provided valuable insights, a larger sample size would be required to ensure statistical power and capture a more comprehensive range of inter-individual variability. While the diversity of participants is a strength, the relatively small number limits the generalizability of some conclusions.

7.5 Conclusion

This study demonstrates that decision support systems have significant potential to affect the scoring accuracy and speed of sleep technologists positively. However, while correct recommendations can make the scoring process more accurate and time-efficient, incorrect recommendations will likely have the opposite effect. Our findings emphasize the critical need for the reliability and correctness of the systems integrated into the workflows of sleep technologists. Additionally, this study provides valuable insights for further research into decision support systems and implementing human-in-the-loop software that incorporates AI into sleep medicine.

Future research should focus on expanding the integration of AI in sleep diagnostics, exploring its application across diverse scoring tasks, and developing systems that empower human experts to deliver more accurate and reliable patient care in less time than currently required. The insights gained from this study pave the way for AI-driven innovations that could revolutionize sleep medicine and enhance patient outcomes. Future research should also examine the long-term

effects of AI integration in the scoring process, specifically whether prolonged exposure to AI influences sleep technologists' scoring behaviors or decision-making habits. Future research could involve longer scoring periods or multiple shorter sessions to address the current limitation of insufficient sleep stage variety, incorporating more diverse and representative sleep segments. Furthermore, given time to adjust and learn to recognize the scoring patterns of the recommendations, the positive effect observed for both scoring accuracy and decision-making time could theoretically increase, however that requires study to confirm. Finally, incorporating information on AI uncertainty can increase the DSS's transparency, increasing the trust and clinical acquiescence of the automatic assistance provided to sleep technologists.

This study underscores the need for the accuracy and reliability of DSS tools when incorporated into the scoring process, as incorrect recommendations were shown to impact sleep technologists' accuracy and decision-making time negatively. Furthermore, the results gathered in this study suggest that take-home PSG solutions do not suffer from reduced accuracy in sleep scoring compared to traditional PSG. However, our findings underscore the risks of over-reliance on AI recommendations, particularly in self-applied PSG when incorrect recommendations are involved. If not carefully managed, such reliance can result in systematic errors, potentially compromising diagnostic reliability. To address these challenges, regular calibration and retraining of AI systems and enhanced training for sleep technologists to effectively collaborate with AI are crucial for mitigating risks and ensuring balanced, reliable outcomes.

Chapter 8

Discussion

"We cannot achieve victory through strength of arms," said Gandalf. "But by endurance, by persistence, and by faith in what lies beyond our sight."

—J.R.R. Tolkien, The Two Towers

This chapter summarizes the key findings of the work comprising the four research chapters in the thesis. It outlines their contributions to further applications in research and practice before covering the limitations of the research conducted as a part of this thesis before finally presenting a forward-looking statement on integrating AI into the workplace.

The first area of research this thesis contributes to is adaptive segmentation within sleep research, with chapter 4 contributing to the design and development of an open-source algorithm designed to locate individual breaths within a respiratory signal. Chapter 4 additionally presents a novel way of analyzing the performance of similar algorithms and provides a rigorous validation of the algorithm. Unlike previous work, which often lacked systematic validation of the proposed algorithms [39], [44], our work includes a rigorous evaluation under multiple conditions. The key takeaways from chapter 4 are the design and implementation of a respiratory cycle isolation algorithm performing at a 94% accuracy. Furthermore, the work concludes by highlighting the need for similar algorithms to be validated

against not only a large amount of data but also data that contains events that might impact the performance or quality of the algorithm's output.

Secondly, this thesis adds to the study of explainable AI in sleep research, particularly in the study of respiration and its mechanics, with chapter 5 describing the training and analysis of a variational autoencoder on respiratory data, utilizing the BreathFinder algorithm to contextualize the respiratory system to its mechanical basis. Existing works focused mainly on applying AI to fixed-length segments [58], [62], while our work leverages the context of the respiratory system for increased model efficiency and explainability. The key takeaways from chapter 5 are that when designing AI models for healthcare applications, immense complexity can be traded off for placing the data in its logical or physiological context (e.g., analyzing respiration based on individual respiratory cycles), it further underlines that the process of respiration is complex and varied, to a degree where even placed in its rightful context, is not trivial for AI to classify events such as obstructive apnea or paradoxical breathing, emphasizing the need to develop AI as a tool in the toolbox of medical experts, rather than a drop-in replacement for the medical professional.

Thirdly, chapter 6 iterates on the contributions of the previous chapters by proposing a robust data ingestion pipeline, which addresses the challenge of creating and processing AI-ready datasets for healthcare, enabling further research. Additionally, chapter 6 demonstrates the challenges of integrating AI into the workflows of sleep technologists. While prior work has highlighted the potential of automated pipelines to streamline clinical workflows [73], challenges remain in scaling these systems for multicentric data ingestion across geographically dispersed centers [126]. Our pipeline bridges this gap, providing not only scalability but also automatic augmentation with AI scoring to reduce manual workload Building on findings by Oxholm et al. [74], our pipeline incorporates analysis into the trust and opinions of sleep technologists to balance automation with oversight, ensuring clinical adoption without compromising trust in the data. The key takeaways of chapter 6 are that to maintain and compose AI-ready datasets, great care must be put into data collection and curation methodologies to ensure the scalability and replicability of research. One such approach is to have a standardized data ingestion pipeline that handles the process and backing up of the data. Furthermore, this chapter outlines the benefits AI can bring in terms of speed and accuracy of sleep technologists when integrated into the scoring process.

Finally, chapter 7 extends further the work in chapter 6 by utilizing novel research methodologies to analyze the scoring habits of sleep technologists when faced with recommendations in the scoring workflows. Unlike prior DSS [72], [92], our research goes further and explores the effects of AI incorrectness, as well as probes for biases for or against human vs. AI recommendations. The key takeaways of chapter 7 are that scoring recommendations in the sleep staging process can positively affect the accuracy of sleep technologists. However, inaccurate rec-

ommendations are likely to result in a loss of accuracy. Furthermore, a key takeaway from the results of chapter 7 is that the participant sleep technologists were not found to display a bias for human recommendations. This finding suggested that sleep technologists can trust automated scoring systems sufficiently for their integration into the scoring process.

8.1 A unified view of AI in the field of sleep research

As presented in chapter 1, this thesis has three main areas of focus, or overarching concepts and contexts that touch on AI in sleep research and medicine, and this thesis is centered around. These areas are covered in the following subsections. Figure 8.1 provides a graphical explanation of the structure of this thesis, with the central circle representing the overarching focus, the three larger circles denoting the main thematic areas, and the smaller circles highlighting the specific contributions within each area.

Figure 8.1: The main focus areas of this thesis and their major contribution towards the central focus of AI in sleep research and medicine.

8.1.1 Designing AI with logical and physiological context

Chapters 4 and 5 demonstrate the goal of creating explainable AI. This involves designing AI that provides clear and intuitive explanations for its decisions rather than relying on black-box decision-making processes. Our results indicate that applying AI to data within its logical context enhances the models' predictive power and makes the model more explainable and intuitive for human experts. These improvements occur without requiring more extensive and complex model sizes or architectures. Explainable AI plays a crucial role in integrating AI into the workflows of sleep researchers and technologists. Transparent AI models enable informed decision-making essential for accurate patient diagnostics.

8.1.2 Standardizing and preparing data pipelines for AI

Projects that include data collection can ensure a standardized and replicable process by designing and implementing AI-enabled data ingestion pipelines such as the one described in chapter 6. Data pipelines allow institutions and projects to ingest more data, using tools that scale efficiently with more data sources without incurring greater workloads for those managing the data. Data pipelines can also increase the consistency and quality of data by integrating features that check for missing or invalid data. If implemented effectively, data pipelines can also enable the integration of diverse data sources into new or existing data sets without incurring the logistical challenges of manually combining heterogeneous data.

8.1.3 Enhancing clinical decision making with human-in-the-loop AI

By integrating AI and DSS directly into the work processes of sleep technologists, numerous benefits can be reaped, such as providing faster diagnostics to patients by accelerating the scoring process, either by selectively focusing the attention of sleep technologists using gray areas or speeding up the decision-making process using AI recommendations. AI as an assistant also has the potential to increase the consistency of sleep staging and the agreement between sleep technologists by providing reliable data-driven recommendations. Our findings also indicate that AI can be a powerful tool in the training and onboarding of new or inexperienced sleep technologists, bridging the gap of experience and helping new sleep technologists enter the workforce smoothly.

8.1.4 Synthesis of the thesis

The areas of focus described in section 8.1.1 and section 8.1.2 contribute to methodologies that promote faster adoption of new AI technologies by ensuring that AI

can be integrated smoothly into the workflows of sleep technologists. Sections 8.1.1 and 8.1.3 add to the literature of increased explainability of AI and means to make the AI outputs actionable by integrating the models directly into the scoring process in the forms of DSS. Finally, sections 8.1.3 and 8.1.2 come together as fully integrated solutions that handle the entire lifecycle of the PSG, from the collection stage to the automatic AI augmentation and ultimately at the hands of the human sleep technologists who remain as the chief decision-makers in the clinical process.

Cumulatively, the chapters paint a larger picture of a top-to-bottom view of the lifecycle of AI in sleep, from designing solutions to curate and prepare data (Chapter 4) to the design of explainable AI (Chapter 5), to describing tools designed to integrate and enable AI in the workflow (Chapter 6) and finally with a more detailed view into how the AI we integrate affect those that work directly with it.

8.2 Towards theoretical applications in the field of integration and application of AI in the clinic

The work described in this thesis has multiple contributions to sleep research, ranging from data collection and preparation to the training and implementation of AI models. Together, chapters 4, 5, 6, and 7 provide a comprehensive framework for integrating AI into sleep research.

Chapter 4 contributes a unique adaptive segmentation algorithm capable of automatically locating individual breaths in a thoracic RIP signal. Segmenting a PSG on a breath-by-breath basis enables varied research into sleep-disordered breathing, where the respiratory system can be analyzed in more detail. Providing a logical segmentation for respiratory data can lead to unique insights into the mechanics of breathing that traditional analyses on fixed-length segments would otherwise miss.

The methodology employed in chapter 5 to preprocess the variable-length respiratory data into an AI-compatible format has the potential for application for other adaptive segmentation such as for EEG signals [27]. Furthermore, our results indicate that contextualizing signal data can increase the explainability of AI models and offload the task of learning the context from the data, allowing the models to focus on capturing the interplay and mechanics of signals.

Chapter 6 provides a framework for large-scale data collection for collaborating researchers and institutions across the earth. Alongside unparalleled collaborative opportunities, the work in Chapter 6 can be extended to serve as a research platform for benchmarking and testing AI models and analyzing their effects on sleep technologists' scoring accuracy and speed.

Chapter 7 boasts a novel technology that can potentially be utilized for further research into sleep and the scoring habits of sleep technologists. Utilizing

remote scoring platforms such as MicroNyx to collect consensus scorings or to employ sleep technologists to provide niche scorings exposes opportunities to analyze sleep on a level of granularity hitherto impossible.

8.3 Towards practical applications in the field of integration and application of AI in the clinic

Our work has made significant contributions to research applications and the practical use of AI in the clinic, ranging from data collection and processing to insights on how AI can successfully be integrated into the workflows of sleep technologists to provide faster and more accurate diagnostics.

Chapter 4 describes an open-source algorithm capable of detecting individual breaths in RIP signals, possibly exposing novel diagnostic workflows for respiratory conditions.

Chapter 5 highlights how contextual AI modeling can improve interpretability and uncover latent patterns in complex data, paving the way for advanced diagnostic tools. Our findings suggest contextualizing AI models can lead to smaller, more cost-efficient models without sacrificing predictive power. Contextualizing data can also improve explainability, enabling more informed decision-making when interpreting the outputs of the AI.

Chapter 6 demonstrates the importance of robust data ingestion pipelines, providing a scalable solution for preparing AI-ready datasets that support diverse research opportunities. Similar to those described in this chapter, such pipelines can provide reliable, automatic pipelines for ingesting, processing, backing up, and scoring PSG in the clinical setting. Additionally, managed processing pipelines enable the adoption and integration of new AI and automated scoring technologies. Integrating AI into the workflows of sleep technologists showed the potential to considerably accelerate the scoring process and improve the scoring agreement between sleep technologists, thus potentially increasing efficiency and reducing workload, thus increasing the capacity for care at healthcare institutions without sacrificing quality or employee well-being. This effect is particularly pronounced for less-experienced sleep technologists, achieving both agreement and scoring times comparable to veteran sleep technologists when using AI assistance.

Finally, the findings from Chapter 7 emphasize the impact of AI-assisted scoring on clinician accuracy, underscoring the importance of designing solutions that enhance the human expert rather than replacing them. Furthermore, remote scoring platforms such as MicroNyx can securely employ sleep technologists worldwide, reducing the need for in-laboratory staff. Our findings suggest that DSS can significantly increase the scoring accuracy of sleep technologists who, according to our findings, have no issue working in settings where AI assists in the scoring

process. Remote scoring platforms can democratize access to expert sleep analysis, particularly for underserved regions.

8.4 Beyond sleep research

The contributions introduced in this work have the potential to impact fields other than sleep research, such as sports science, where research into respiration can provide research opportunities on the link between respiration and performance [173], as well as enable efficient real-time monitoring of respiration in non-clinical settings. Standardized pipelines can streamline the integration of remotely collected data in telemedicine, ensuring consistency in settings where clinical oversight is impossible [174]. Data collection is a task not unique to sleep research, and projects in various fields across academia can gain valuable insights on data collection and AI integration from chapter 6. Studies into the EEG signal in fields outside of sleep research and medicine may also find value in the contributions, particularly in chapter 7, where remote platforms can be utilized to employ experts worldwide for data analysis and manual annotation or remote monitoring [175]. Logical segmentation, or adaptive segmentation, was initially proposed for EEG data [27] and inspired much of the work done in this thesis. Still, adaptive segmentation could have broader applications than sleep research, such as for ECG or EEG analysis in studies of neurological or cardiovascular disorders.

8.5 Limitations

Although this work represents a collection of vital contributions towards AI in sleep medicine, it has some notable limitations. The first and most prominent limitation is that chapters 6 and 7 suffer from a low sample size of sleep technologists participating in the data collection. Unfortunately, this is a possible symptom of the high workloads of sleep technologists, who cannot find time in their busy schedules to participate in time-consuming research. A step towards mitigating this limitation in future works is offering compensation, longer participation windows, or more comprehensive advertising, to name a few strategies.

Another limitation is that while chapter 7 provides a comprehensive analysis of scoring patterns, the limited amount of signal data used for the study does not give a generalizable view of the scoring patterns, whereas a longer scoring window might provide insights into phenomena such as scoring fatigue. The relatively short hour-scoring period was a design decision for this work. However, future studies could repeat the experiment with extended scoring periods, using dedicated staff as participants. However, longer scoring periods may necessarily lead to lower participant numbers due to the high workloads of sleep technologists.

While chapter 5 advances the literature on the explainability of AI in sleep research, the overlap of events across clusters was significant, potentially reducing explainability due to the substantial spread of respiratory events across clusters. A possible solution to this limitation is training a classifier in parallel to assign labels such as apnea or movement events to the breaths in the latent space of the VAE, including the accuracy of the classifier in the loss function of the VAE itself, thus forcing the VAE to learn to separate the breaths more efficiently in the latent space. This method was not employed for this work since the aim was to provide a view into how AI interpreted and understood the respiratory system. In its reconstructions, the VAE tended not to include the higher-frequency components of the input signals in its output. This high-frequency data may theoretically include physiological or pathophysiological information but was not represented in the VAE output either due to the VAE learning to ignore it or due to the deisgn limitations of the model.

Although this thesis gives a comprehensive view of the integration of AI into sleep research, it does not cover the validation and verification of the AI models themselves, a vital part of ensuring correctness and thus increasing the likelihood of reaping the positive benefits of AI highlighted in this work. Addressing any of the limitations in this section in future studies could provide a more comprehensive and generalizable understanding of scoring patterns and fatigue, ultimately leading to better integration of AI into clinical workflows.

Chapter 9

Conclusion

"I am glad you are here with me. Here at the end of all things, Sam."

—J.R.R. Tolkien, The Return of the King

This thesis covers the work and contributions to the field of study of integrating AI in sleep research and medicine. Integrating AI into any healthcare setting faces many challenges, such as heterogeneity of data, lack of transparency of the AI, high requirements for trust from patients and healthcare professionals, and logistical issues with its integration. Lack of explainability of AI solutions can negatively affect the trust of sleep technologists in them [23], and a lack of trust in the AI may lead to the sleep technologists refusing to engage with it, negating its potential benefit [78]. This thesis tackles these problems by approaching them from three main areas of focus:

1. Increasing AI explainability by designing models to analyze signal data in logically physiological contexts.
2. Creating standardized and scalable AI-enabled data pipelines.
3. Enhancing clinical workflows with human-in-the-loop AI solutions.

Chapter 1 introduced the background and clarified the context this thesis exists in. Chapter 2 explored related work and positions this thesis amongst the existing body of research. Chapters 4 through 7 contain this thesis's four publications. Chapter 8 outlines the main findings of the publications and relates them to the research questions introduced in chapter 1. This chapter seeks to conclude the thesis, presenting the main contributions and concluding in a forward-looking statement.

This thesis takes a holistic approach to AI integration in sleep, spanning algorithm design, data processing, model design, and clinical integration. The research conducted as part of this thesis paves the way for broader AI adoption in healthcare by providing tools to handle data from various stages of the PSG lifecycle. Our findings demonstrated improvements in efficiency, trust and accuracy from integrating AI tools into the workflows of sleep technologists.

9.1 Main Contributions

This thesis presents contributions toward facilitating the adoption of AI in sleep medicine by designing solutions that address challenges in data pipelines and enhance clinical decision-making. These goals were formulated as the three main questions, first outlined in section 1.1. Each question is addressed through specific chapters, outlining the contributions to advancing AI adoption in sleep medicine.

- **RQ1:** *What could a fully-managed PSG ecosystem look like?* Addressed in Chapters 4, 6, and 7, with specific emphasis on the AI-enabled data ingestion pipelines and decision support DSS. Together, these contributions outline the core components of a fully managed PSG ecosystem.
- **RQ2:** *How does one successfully integrate AI into sleep research or sleep medicine workflows?* Contributed to by all chapters, particularly Chapters 5 and 7, by demonstrating the importance of explainability, intuitiveness, and the integration of human expertise. From interpretable AI models (Chapter 5) to the role of DSS in keeping humans as the final decision-makers (Chapter 7), these works provide insights into successful AI integration.
- **RQ3:** *What are the effects of integrating AI into the scoring process?* Explored in detail in Chapters 6 and 7, focusing on the significant improvements in scoring accuracy and efficiency brought by AI augmentation and the challenges posed by incorrect recommendations.

Our findings from chapters 4, 5, 6, and 7 have allowed us to address these questions, either on their own or collaboratively. In Chapter 4, we discuss the design

and development of a novel respiratory cycle isolation algorithm and carefully validate it against a set of hand-labeled respiratory cycles. The algorithm performs excellently, achieving 94% accuracy on the hand-labeled data. Our findings provide significant contributions to the field of adaptive segmentation in respiratory data. For AI to be responsibly integrated into the clinician or researcher workflows, a degree of explainability and intuitiveness is required. Algorithms that segment data into its logical contexts have the potential to increase the interpretability and intuitiveness of algorithms that are trained on or make decisions on the basis of that context. This work addresses RQ1 since, as our findings from chapter 5 suggest, automatic methods to contextualize data are vital to include in PSG ecosystems that intend to incorporate explainable AI into the workflow.

Chapter 5 covers our work in implementing and training a VAE model on respiratory data extracted and prepared using the BreathFinder algorithm introduced in chapter 4. Despite significant compression, we found that the VAE could still broadly accurately reconstruct the respiratory cycles, despite filtering out some of the high-frequency elements of the signals. Furthermore, we found that respiratory events clustered together in the latent space to a varying degree. This work provides a glimpse into the understanding of the respiratory system on a breath-by-breath basis, as seen by an unsupervised AI model. The analysis and study of such models can potentially improve our understanding of these systems and increase AI explainability. This work addresses RQ2, as explainability and intuitiveness are paramount to successfully integrate AI into the clinical process.

Chapter 6 describes the work done in the implementation and analysis of an AI-enabled data ingestion pipeline and analysis of how AI augmentation affects the scoring process of sleep technologists. In addition, we propose a new term for the willingness to accept AI into the workflows of sleep technologists, 'clinical acuity.' We found that AI augmentation has the potential to shorten the scoring process by up to 65 minutes in one case. This work represents an important contribution toward the basis of a fully managed PSG ecosystem by enabling consistent and reproducible data ingestion and processing. This work addresses RQ1 as data ingestion pipelines are the logical entry points for PSG data into the ecosystem. RQ2 is also answered in this work, as successfully integrating AI into the sleep research or medicine workflows demands the consistent linear processes pipelines can provide. Finally, this work touches on RQ3, as we find from our analysis that AI in the workflow can significantly speed up scoring time and improve accuracy.

Chapter 7 delves into our research on DSS as a scoring aide. Similarly to our work in 6, we analyzed the effects of including recommendations in the scoring system, except to a much more granular level than in the previous work. In addition, we analyzed both the effects recommendations had when presented by a human vs. an AI and the difference between correct and incorrect recommendations. Our findings indicate that correct recommendations can have a significant positive effect on scoring accuracy and decision-making time, while incorrect rec-

ommendations have a significant adverse effect on scoring accuracy. Our findings indicated no inherent bias for or against human recommendations over AI ones, indicating a high level of clinical acuity towards AI in the scoring process. This work contributes towards a fuller image of the fully-managed PSG ecosystem by presenting the PSG to human experts for final decision-making, including the human back in the scoring loop. This work addresses all three research questions, as DSS can act as a bridge between the AI and the human experts, serving as the last link of a chain that starts with the data ingestion pipelines covered in chapter 6 (RQ1), Furthermore, we conclude in chapter 7 that to successfully integrate AI into the sleep scoring process, the human expert must be kept in the loop as the final decision-maker. Finally, our findings suggest that AI in the role of a DSS can significantly improve the scoring accuracy of sleep technologists (RQ3).

9.2 Future work

Future work enabled or exposed by this work is varied, touching on many aspects of sleep research and medicine.

Adaptive segmentation approaches have the strength of leveraging the physiological phases of various biological signals to segment data into logical periods. While this work mainly targeted respiratory signals, a wide variety of contexts are liable to gain from methodologies similar to those applied in chapter 4 such as research into muscle activity on the EMG signals, heart activity on the ECG signal, or apnea research on desaturations in the oxygen signals to name a few.

Our findings in chapter 5 suggest that further research into the respiratory system on a breath-by-breath basis is warranted. Furthermore, we believe that further research into the application of AI on adaptively segmented data may reveal opportunities for simpler and more explainable model designs for a diverse range of signal data such as EEG, EMG, ECG, or other respiratory data.

Data platforms in sleep research and medicine are powerful tools with wide-ranging potential. Our findings suggest that data platforms can serve an essential role in the collection and processing of data. Future endeavors towards data platforms in healthcare may find opportunities in more diverse data collection scenarios, such as recording fitness evaluations, collecting brainwave data for cognitive research, or ingesting and integrating genome data to name a few possibilities. Research into data platforms and pipelines also enables further research into the automatic integration of AI.

Our findings indicate that DSS can greatly influence scoring time and accuracy. Subsequent research opportunities this work exposes could include active learning applications where the human expert corrections are used to train the model in an active learning approach. Other avenues of investigation exposed are further

investigations into how explanations of the AI's decision-making affect the clinical acquiescence of the models in the workflow.

Finally, it is the opinion of the authors of this thesis that any future attempt to integrate AI into sleep medicine, or any field within healthcare, should have as its primary goal not to replace the human element but rather to uplift it.

Bibliography

[1] E. Brynjolfsson and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, 1st. W. W. Norton & Company, 2014, ISBN: 0393239357.

[2] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp. 42–47. DOI: 10.1109/CTS.2013.6567202.

[3] M. Kryger, T. Roth, and W. C. Dement, Eds., *Principles and Practice of Sleep Medicine*, 6th. Philadelphia, PA: Elsevier, 2017, ISBN: 978-0-323-24288-2. DOI: 10.1016/C2012-0-03543-0.

[4] E. S. Arnardottir, E. Bjornsdottir, K. A. Olafsdottir, B. Benediktsdottir, and T. Gislason, "Obstructive sleep apnoea in the general population: Highly prevalent but minimal symptoms," *European Respiratory Journal*, vol. 47, no. 1, pp. 194–202, 2016. DOI: 10.1183/13993003.01148-2015.

[5] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis," *Lancet Respir Med*, vol. 7, no. 8, pp. 687–698, Aug. 2019.

[6] N. F. Watson, "Health care savings: The economic value of diagnostic and therapeutic care for obstructive sleep apnea," en, *J. Clin. Sleep Med.*, vol. 12, no. 8, pp. 1075–1077, Aug. 2016.

[7] M. Troester, S. Quan, B. RB., *et al.*, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* Darien, IL: American Academy of Sleep Medicine, 2023, vol. Version 3.0.

[8] American Academy of Sleep Medicine, "International classification of sleep disorders, third edition," 2014.

[9] S. L. Himanen and J. Hasan, "Limitations of rechtschaffen and kales," en, *Sleep Med. Rev.*, vol. 4, no. 2, pp. 149–167, Apr. 2000.

[10] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," en, *J. Clin. Sleep Med.*, vol. 9, no. 1, pp. 81–87, Jan. 2013.

[11] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Respiratory events," en, *J. Clin. Sleep Med.*, vol. 10, no. 4, pp. 447–454, Apr. 2014.

[12] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd. USA: Prentice Hall Press, 2009, ISBN: 0136042597.

[13] D. Coyle and L. Hampton, "21st century progress in computing," *Telecomm. Policy*, vol. 48, no. 1, p. 102 649, Feb. 2024.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," en, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[15] T. M. Mitchell, *Machine Learning*. New York, NY: McGraw-Hill, 1997, vol. 1.

[16] J. M. Faria, "Non-determinism and failure modes in machine learning," in *2017 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, IEEE, Oct. 2017, pp. 310–316.

[17] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:57574615.

[18] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of explainable ai techniques in healthcare," *Sensors*, vol. 23, no. 2, 2023, ISSN: 1424-8220. DOI: 10.3390/s23020634. [Online]. Available: https://www.mdpi.com/1424-8220/23/2/634.

[19] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," en, *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019.

[20] H. Smith, "Clinical AI: Opacity, accountability, responsibility and liability," *AI Soc.*, vol. 36, no. 2, pp. 535–545, Jun. 2021.

[21] P. Kumar, S. Chauhan, and L. K. Awasthi, "Artificial intelligence in healthcare: Review, ethics, trust challenges & future research directions," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105 894, 2023, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2023.105894. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623000787.

[22] S. Kumar, S. Datta, V. Singh, D. Datta, S. K. Singh, and R. Sharma, "Applications, challenges, and future directions of Human-in-the-Loop learning," *IEEE Access*, vol. 12, pp. 75 735–75 760, 2024.

[23] Nzenwata, Ilori, Tai-Ojuolape, *et al.*, "Explainable AI: A systematic literature review focusing on healthcare," en, *J. Comput. Sci. Appl.*, vol. 12, no. 1, pp. 10–16, Aug. 2024.

[24] A. Rayan, A. B. Szabo, and L. Genzel, "The pros and cons of using automated sleep scoring in sleep research: Comparative analysis of automated sleep scoring in human and rodents: advantages and limitations," *Sleep*, zsad275, Oct. 2023, ISSN: 0161-8105. DOI: 10.1093/sleep/zsad275. eprint: https://academic.oup.com/sleep/advance-article-pdf/doi/10.1093/sleep/zsad275/53670246/zsad275.pdf. [Online]. Available: https://doi.org/10.1093/sleep/zsad275.

[25] L. Fiorillo, A. Puiatti, M. Papandrea, *et al.*, "Automated sleep scoring: A review of the latest approaches," en, *Sleep Med. Rev.*, vol. 48, p. 101 204, Dec. 2019.

[26] F. Mendonça, S. S. Mostafa, A. G. Ravelo-García, F. Morgado-Dias, and T. Penzel, "A review of obstructive sleep apnea detection approaches," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 825–837, 2019. DOI: 10.1109/JBHI.2018.2823265.

[27] H. M. Praetorius, G. Bodenstein, and O. D. Creutzfeldt, "Adaptive segmentation of EEG records: A new approach to automatic EEG analysis," en, *Electroencephalogr. Clin. Neurophysiol.*, vol. 42, no. 1, pp. 84–94, Jan. 1977.

[28] E. LaRosa and D. Danks, "Impacts on trust of healthcare ai," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18, New Orleans, LA, USA: Association for Computing Machinery, 2018, pp. 210–215, ISBN: 9781450360128. DOI: 10.1145/3278721.3278771. [Online]. Available: https://doi.org/10.1145/3278721.3278771.

[29] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," en, *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[30] A. Ermshaus, P. Schäfer, and U. Leser, "ClaSP: Parameter-free time series segmentation," en, *Data Min. Knowl. Discov.*, vol. 37, no. 3, pp. 1262–1300, May 2023.

[31] G. Bodenstein, W. Schneider, and C. V. Malsburg, "Computerized EEG pattern classification by adaptive segmentation and probability-density-function classification. description of the method," en, *Comput. Biol. Med.*, vol. 15, no. 5, pp. 297–313, 1985.

[32] R. Agarwal and J. Gotman, "Adaptive segmentation of electroencephalographic data using a nonlinear energy operator," vol. 4, 199–202 vol.4, May 1999.

[33] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of Inter-Scorer variability," en, *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, Jun. 2016.

[34] M. K. Sein, O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren, "Action design research," *MIS Quarterly*, vol. 35, no. 1, pp. 37–56, 2011, ISSN: 02767783. [Online]. Available: http://www.jstor.org/stable/23043488 (visited on 12/30/2024).

[35] C. Herzog, S. Blank, and B. C. Stahl, "Towards trustworthy medical AI ecosystems – a proposal for supporting responsible innovation practices in AI-based medical innovation," en, *AI Soc.*, pp. 1–21, Oct. 2024.

[36] M. Findlay and J. Seah, "An ecosystem approach to ethical AI and data use: Experimental reflections," en, in *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*, IEEE, Sep. 2020, pp. 192–197.

[37] T. P. Moyles, R. F. Erlandson, and T. Roth, "A nonparametric statistical approach to breath segmentation," in *Images of the Twenty-First Century. Proceedings of the Annual International Engineering in Medicine and Biology Society,*, Nov. 1989, 330–331 vol.1. DOI: 10.1109/IEMBS.1989.95756.

[38] R. D. Chervin, J. W. Burns, N. S. Subotic, C. Roussi, B. Thelen, and D. L. Ruzicka, "Method for detection of respiratory cycle-related EEG changes in sleep-disordered breathing," *Sleep*, vol. 27, no. 1, pp. 110–115, Feb. 2004.

[39] P. Lopez-Meyer and E. Sazonov, "Automatic breathing segmentation from wearable respiration sensors," in *2011 Fifth International Conference on Sensing Technology*, Nov. 2011, pp. 156–160. DOI: 10.1109/ICSensT.2011.6136953.

[40] T. Rosenwein, E. Dafna, A. Tarasiuk, and Y. Zigel, "Detection of breathing sounds during sleep using non-contact audio recordings," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2014, pp. 1489–1492, 2014.

[41] O. Yahya and M. Faezipour, "Automatic detection and classification of acoustic breathing cycles," in *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*, 2014, pp. 1–5. DOI: 10.1109/ASEEZone1.2014.6820648.

[42] M. Włodarczak, "RespInPeace: Toolkit for processing respiratory belt data," Jun. 2019. DOI: 10.5281/ZENODO.3246019. [Online]. Available: https://zenodo.org/record/3246019 (visited on 08/05/2021).

[43] P. Hult, B. Wranne, and P. Ask, "A bioacoustic method for timing of the different phases of the breathing cycle and monitoring of breathing frequency," *Med. Eng. Phys.*, vol. 22, no. 6, pp. 425–433, Jul. 2000.

[44] P. Hult, T. Fjällbrant, B. Wranne, O. Engdahl, and P. Ask, "An improved bioacoustic method for monitoring of respiration," *THC*, vol. 12, no. 4, pp. 323–332, Oct. 2004.

[45] C.-H. Hsiao, T.-W. Lin, C.-W. Lin, *et al.*, "Breathing sound segmentation and detection using transfer learning techniques on an Attention-Based Encoder-Decoder architecture," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2020, pp. 754–759.

[46] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "Adaptive neuro-fuzzy inference system for breath phase detection and breath cycle segmentation," *Comput. Methods Programs Biomed.*, vol. 145, pp. 67–72, Jul. 2017.

[47] W. Lalouani, M. Younis, R. N. Emokpae Jr, and L. E. Emokpae, "Enabling effective breathing sound analysis for automated diagnosis of lung diseases," en, *Smart Health (Amst)*, vol. 26, p. 100 329, Dec. 2022.

[48] H. Alshaer, G. R. Fernie, E. Sejdić, and T. D. Bradley, "Adaptive segmentation and normalization of breathing acoustic data of subjects with obstructive sleep apnea," in *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, 2009, pp. 279–284. DOI: 10.1109/TIC-STH.2009.5444489.

[49] J. Korten and G. Haddad, "Respiratory waveform pattern recognition using digital techniques," *Computers in Biology and Medicine*, vol. 19, no. 4, pp. 207–217, 1989, ISSN: 0010-4825. DOI: https://doi.org/10.1016/0010-4825(89)90009-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0010482589900097.

[50] H. Korkalainen, J. Aakko, S. Nikkonen, *et al.*, "Accurate deep Learning-Based sleep staging in a clinical population with suspected obstructive sleep apnea," en, *IEEE J Biomed Health Inform*, vol. 24, no. 7, pp. 2073–2081, Jul. 2020.

[51] L. Cen, Z. L. Yu, T. Kluge, and W. Ser, "Automatic system for obstructive sleep apnea events detection using convolutional neural network," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 3975–3978.

[52] A. Thommandram, J. M. Eklund, and C. McGregor, "Detection of apnoea from respiratory time series data using clinically recognizable features and kNN classification," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2013, pp. 5013–5016.

[53] T. Rosenwein, E. Dafna, A. Tarasiuk, and Y. Zigel, "Breath-by-breath detection of apneic events for OSA severity estimation using non-contact audio recordings," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 7688–7691.

[54] S. Nikkonen, H. Korkalainen, A. Leino, *et al.*, "Automatic respiratory event scoring in obstructive sleep apnea using a long Short-Term memory neural network," en, *IEEE J Biomed Health Inform*, vol. 25, no. 8, pp. 2917–2927, Aug. 2021.

[55] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.

[56] W. Xu, H. Sun, C. Deng, and Y. Tan, "Variational autoencoder for Semi-Supervised text classification," in *Thirty-First AAAI Conference on Artificial Intelligence*, Feb. 2017.

[57] X. Li and J. She, "Collaborative variational autoencoder for recommender systems," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Halifax, NS, Canada: Association for Computing Machinery, Aug. 2017, pp. 305–314.

[58] N. Costa, L. Sánchez, and I. Couso, "Semi-Supervised recurrent variational autoencoder approach for visual diagnosis of atrial fibrillation," *IEEE Access*, vol. 9, pp. 40 227–40 239, 2021.

[59] V. V. Kuznetsov, V. A. Moskalenko, D. V. Gribanov, and N. Y. Zolotykh, "Interpretable feature generation in ECG using a variational autoencoder," en, *Front. Genet.*, vol. 12, p. 638 191, Apr. 2021.

[60] A. Oluwasanmi, M. U. Aftab, E. Baagyere, Z. Qin, M. Ahmad, and M. Mazzara, "Attention autoencoder for generative latent representational learning in anomaly detection," en, *Sensors*, vol. 22, no. 1, Dec. 2021.

[61] A. Singh and T. Ogunfunmi, "An overview of variational autoencoders for source separation, finance, and Bio-Signal applications," en, *Entropy*, vol. 24, no. 1, Dec. 2021.

[62] O. Pastor-Serrano, D. Lathouwers, and Z. Perkó, "A semi-supervised autoencoder framework for joint generation and classification of breathing," en, *Comput. Methods Programs Biomed.*, vol. 209, p. 106 312, Sep. 2021.

[63] M. de Reuver, C. Sørensen, and R. C. Basole, "The digital platform: A research agenda," *Journal of Information Technology*, vol. 33, no. 2, pp. 124–135, 2018. DOI: 10.1057/s41265-016-0033-3. eprint: https://doi.org/10.1057/s41265-016-0033-3. [Online]. Available: https://doi.org/10.1057/s41265-016-0033-3.

[64] A. Gleiss, M. Kohlhagen, and K. Pousttchi, "An apple a day - how the platform economy impacts value creation in the healthcare market," en, *Electron. Mark.*, vol. 31, no. 4, pp. 849–876, Apr. 2021.

[65] S. Hermes, T. Riasanow, E. K. Clemons, M. Böhm, and H. Krcmar, "The digital transformation of the healthcare industry: Exploring the rise of emerging platform ecosystems and their influence on the role of patients," en, *Bus. Res.*, vol. 13, no. 3, pp. 1033–1069, Nov. 2020.

[66] A. Newman, Y. L. Bavik, M. Mount, and B. Shao, "Data collection via online platforms: Challenges and recommendations for future research," en, *Appl. Psychol.*, vol. 70, no. 3, pp. 1380–1402, Jul. 2021.

[67] K. Cresswell, A. Majeed, D. Bates, and A. Sheikh, "Computerised decision support systems for healthcare professionals: An interpretative review," English, *Informatics in Primary Care*, vol. 20, no. 2, pp. 115–128, Feb. 2012, ISSN: 1476-0320. DOI: 10.14236/jhi.v20i2.32.

[68] B. F. Sveinbjarnarson, L. Schmitz, E. S. Arnardottir, and A. S. Islind, "The sleep revolution platform: A dynamic data source pipeline and digital platform architecture for complex sleep data," en, *Curr. Sleep Med. Rep.*, vol. 9, no. 2, pp. 91–100, Apr. 2023.

[69] S. Garbarino and N. L. Bragazzi, "Revolutionizing sleep health: The emergence and impact of personalized sleep medicine," en, *J. Pers. Med.*, vol. 14, no. 6, p. 598, Jun. 2024.

[70] M. P. Fanti and W. Ukovich, "Decision support systems and integrated platforms: New approaches for managing systems of systems," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 3, no. 1, pp. 18–22, 2017. DOI: 10.1109/MSMC.2016.2623868.

[71] D. Alvarez-Estevez and R. M. Rijsman, "Computer-assisted analysis of polysomnographic recordings improves inter-scorer associated agreement and scoring times," *PLoS One*, vol. 17, no. 9, e0275530, Sep. 2022.

[72] S.-F. Liang, Y.-H. Shih, P.-Y. Chen, and C.-E. Kuo, "Development of a human-computer collaborative sleep scoring system for polysomnography recordings," *PLoS One*, vol. 14, no. 7, e0218948, Jul. 2019.

[73] B. P. Choo, Y. Mok, H. C. Oh, *et al.*, "Benchmarking performance of an automatic polysomnography scoring system in a population with suspected sleep disorders," *Frontiers in Neurology*, vol. 14, 2023, ISSN: 1664-2295. DOI: 10.3389/fneur.2023.1123935. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fneur.2023.1123935.

[74] C. Oxholm, A.-M. S. Christensen, R. Christiansen, U. K. Wiil, and A. S. Nielsen, "Attitudes of patients and health professionals regarding screening algorithms: Qualitative study," *JMIR Form Res*, vol. 5, no. 8, e17971, Aug. 2021, ISSN: 2561-326X. DOI: 10.2196/17971.

[75] V. Gerla, V. Kremen, M. Macas, E. Saifutdinova, A. Mladek, and L. Lhotska, "Expert-in-the-loop learning for sleep EEG data," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2018, pp. 2590–2596.

[76] V. Gerla, V. Kremen, M. Macas, *et al.*, "Iterative expert-in-the-loop classification of sleep PSG recordings using a hierarchical clustering," *J. Neurosci. Methods*, vol. 317, pp. 61–70, Apr. 2019.

[77] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 624–635, ISBN: 9781450383097. DOI: 10.1145/3442188.3445923. [Online]. Available: https://doi.org/10.1145/3442188.3445923.

[78] P. D. Hyesun Choung and A. Ross, "Trust in ai and its role in the acceptance of ai technologies," *International Journal of Human–Computer Interaction*, vol. 39, no. 9, pp. 1727–1739, 2023. DOI: 10.1080/10447318.2022.2050543. eprint: https://doi.org/10.1080/10447318.2022.2050543. [Online]. Available: https://doi.org/10.1080/10447318.2022.2050543.

[79] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, and M. L. Tielman, "A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges," *ACM J. Responsib. Comput.*, vol. 1, no. 4, Nov. 2024. DOI: 10.1145/3696449. [Online]. Available: https://doi.org/10.1145/3696449.

[80] T. P. Quinn, M. Senadeera, S. Jacobs, S. Coghlan, and V. Le, "Trust and medical ai: The challenges we face and the expertise needed to overcome them," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 890–894, Dec. 2020, ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa268. eprint: https://academic.oup.com/jamia/article-pdf/28/4/890/36642134/ocaa268.pdf. [Online]. Available: https://doi.org/10.1093/jamia/ocaa268.

[81] A. Klingbeil, C. Grützner, and P. Schreck, "Trust and reliance on AI — an experimental study on the extent and costs of overreliance on AI," en, *Comput. Human Behav.*, vol. 160, no. 108352, p. 108 352, Nov. 2024.

[82] N. Leveson and C. Turner, "An investigation of the therac-25 accidents," *Computer*, vol. 26, no. 7, pp. 18–41, 1993. DOI: 10.1109/MC.1993.274940.

[83] S. Passi and M. Vorvoreanu, "Overreliance on ai literature review," *Microsoft Research*, 2022.

[84] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna, "Explanations can reduce overreliance on ai systems during decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW1, Apr. 2023. DOI: 10.1145/3579605. [Online]. Available: https://doi.org/10.1145/3579605.

[85] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, Apr. 2021. DOI: 10.1145/3449287. [Online]. Available: https://doi.org/10.1145/3449287.

[86] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review," en, *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022.

[87] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable q-factor wavelet transform and spectral features," en, *J. Neurosci. Methods*, vol. 271, pp. 107–118, Sep. 2016.

[88] A. R. Hassan and A. Subasi, "A decision support system for automated identification of sleep stages from single-channel EEG signals," *Knowledge-Based Systems*, vol. 128, pp. 115–124, Jul. 2017.

[89] S. Charbonnier, L. Zoubek, S. Lesecq, and F. Chapotot, "Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging," en, *Comput. Biol. Med.*, vol. 41, no. 6, pp. 380–389, Jun. 2011.

[90] D. Kim, J. Lee, Y. Woo, J. Jeong, C. Kim, and D.-K. Kim, "Deep learning application to clinical decision support system in sleep stage classification," en, *J Pers Med*, vol. 12, no. 2, Jan. 2022.

[91] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-sleep: Resilient high-frequency sleep staging," *NPJ digital medicine*, vol. 4, no. 1, p. 72, 2021.

[92] J. Hwang, T. Lee, H. Lee, and S. Byun, "A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: User-Centered design and evaluation study," en, *J. Med. Internet Res.*, vol. 24, no. 1, e28659, Jan. 2022.

[93] M. Serwatko, "Validation of a new method to assess respiratory effort non-invasively," M.S. thesis, Reykjavik University, 2016. eprint: http://hdl.handle.net/1946/23804.

[94] A. F. Horne, K. A. Olafsdottir, and E. S. Arnardottir, "In-person versus video hookup instructions: A comparison of home sleep apnea testing quality," *Journal of Clinical Sleep Medicine*, Apr. 2022.

[95] G. Jouan, E. S. Arnardottir, A. S. Islind, and M. Óskarsdóttir, "An algorithmic approach to identification of gray areas: Analysis of sleep scoring expert ensemble non agreement areas using a multinomial mixture model," *European Journal of Operational Research*, 2023.

[96] M. Rusanen, G. Jouan, R. Huttunen, *et al.*, "Asaga: Automatic sleep analysis with gray areas," 2023. arXiv: 2310.02032 [cs.LG].

[97] M. Rusanen, H. Korkalainen, H. Gretarsdottir, *et al.*, "Self-applied somnography: Technical feasibility of electroencephalography and electro-oculography signal characteristics in sleep staging of suspected sleep-disordered adults," en, *J. Sleep Res.*, vol. 33, no. 2, e13977, Apr. 2024.

[98] D. Roselli, J. Matthews, and N. Talagala, "Managing bias in ai," in *Companion Proceedings of The 2019 World Wide Web Conference*, ser. WWW '19, San Francisco, USA: Association for Computing Machinery, 2019, pp. 539–544, ISBN: 9781450366755. DOI: 10.1145/3308560.3317590. [Online]. Available: https://doi.org/10.1145/3308560.3317590.

[99] F. Portier, A. Portmann, P. Czernichow, *et al.*, "Evaluation of home versus laboratory polysomnography in the diagnosis of sleep apnea syndrome," *American Journal of Respiratory and Critical Care Medicine*, vol. 162, no. 3, pp. 814–818, 2000, PMID: 10988088. DOI: 10.1164/ajrccm.162.3.9908002. eprint: https://doi.org/10.1164/ajrccm.162.3.9908002. [Online]. Available: https://doi.org/10.1164/ajrccm.162.3.9908002.

[100] H. Schulz, "Rethinking sleep analysis," en, *J. Clin. Sleep Med.*, vol. 4, no. 2, pp. 99–103, Apr. 2008.

[101] A. Procházka, J. Kuchyňka, M. Yadollahi, C. P. S. Araujo, and O. Vyšata, "Adaptive segmentation of multimodal polysomnography data for sleep stages detection," in *2017 22nd International Conference on Digital Signal Processing (DSP)*, Aug. 2017, pp. 1–4.

[102] H. Koch, P. Jennum, and J. A. E. Christensen, "Automatic sleep classification using adaptive segmentation reveals an increased number of rapid eye movement sleep transitions," en, *J. Sleep Res.*, vol. 28, no. 2, e12780, Apr. 2019.

[103] B. H. Þórðarson, "Analysis and detection of obstructive apnea in individual breath cycles," en, M.S. thesis, Reykjavik University, Jun. 2021.

[104] A. S. BaHammam, "Signal failure of type 2 comprehensive unattended sleep studies in patients with suspected respiratory sleep disordered breathing," *Sleep Breath.*, vol. 9, no. 1, pp. 7–11, Mar. 2005.

[105] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[106] M. Vlachos, P. Yu, and V. Castelli, "On periodicity detection and structural periodic similarity," in *Proceedings of the 2005 SIAM International Conference on Data Mining*. 2005, pp. 449–460. DOI: 10.1137/1.9781611972757.40.

[107] B. Thordarson, A. S. Islind, E. Arnardottir, and M. Óskarsdóttir, "Exploration of sleep events in the latent space of variational autoencoders on a Breath-by-Breath basis," en, in *Proceedings of the 56th Hawaii International Conference on System Sciences*, Maui, Hawaii: Computer Society Press, pp. 3091–30 911.

[108] E. S. Arnardottir, A. S. Islind, and M. Óskarsdóttir, "The future of sleep measurements: A review and perspective," *Sleep medicine clinics*, vol. 16, no. 3, pp. 447–464, 2021.

[109] A. Malhotra, M. Younes, S. T. Kuna, *et al.*, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," en, *Sleep*, vol. 36, no. 4, pp. 573–582, Apr. 2013.

[110] I. Perez-Pozuelo, B. Zhai, J. Palotti, *et al.*, "The future of sleep health: A data-driven revolution in sleep science and medicine," en, *NPJ Digit Med*, vol. 3, p. 42, Mar. 2020.

[111] E. Mekov, M. Miravitlles, and R. Petkov, "Artificial intelligence and machine learning in respiratory medicine," en, *Expert Rev. Respir. Med.*, vol. 14, no. 6, pp. 559–564, Jun. 2020.

[112] A. Bandyopadhyay and C. Goldstein, "Clinical applications of artificial intelligence in sleep medicine: A sleep clinician's perspective," *Sleep and Breathing*, Mar. 2022.

[113] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," en, *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[114] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2014. arXiv: 1312.6114 [stat.ML].

[115] J. Gerlings, A. Shollo, and I. Constantiou, "Reviewing the need for explainable artificial intelligence (xAI)," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, Hawaii International Conference on System Sciences, 2021.

[116] D. Charte, F. Charte, M. J. del Jesus, and F. Herrera, "An analysis on the use of autoencoders for representation learning: Fundamentals, learning task case studies, explainability and challenges," *Neurocomputing*, vol. 404, pp. 93–107, Sep. 2020.

[117] B. Holm, M. ÓskarsdÓttir, E. S. ArnardÓttir, M. Serwatko, J. Mallett, and M. Borsky, "Automatic Non-Invasive isolation of respiratory cycles," *arXiv*, Mar. 2022. arXiv: 2203.01828 [physics.med-ph].

[118] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[119] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, and Z. C. et. al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015.

[120] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[121] K. W. De Bock, K. Coussement, A. De Caigny, *et al.*, "Explainable ai for operational research: A defining framework, methods, applications, and a research agenda," *European Journal of Operational Research*, 2023.

[122] E. Jermutus, D. Kneale, J. Thomas, and S. Michie, "Influences on user trust in healthcare artificial intelligence: A systematic review," *Wellcome Open Research*, vol. 7, 2022.

[123] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, p. 107 161, 2022.

[124] R. S. Jang, D. Ciliberti, E. A. Mankin, and G. R. Poe, "Recurrent hippocampo-neocortical sleep-state divergence in humans," *Proceedings of the National Academy of Sciences*, vol. 119, no. 44, e2123427119, 2022.

[125] D. A. Pevernagie and E. S. Arnardottir, "Looking for clues in the hypnogram - the human eye and the machine," *Sleep*, Jan. 2024.

[126] E. S. Arnardottir, A. S. Islind, M. Óskarsdóttir, *et al.*, "The sleep revolution project: The concept and objectives," en, *J. Sleep Res.*, vol. 31, no. 4, e13630, Aug. 2022.

[127] L. Biedebach, M. Óskarsdóttir, E. S. Arnardóttir, *et al.*, "Anomaly detection in sleep: Detecting mouth breathing in children," *Data Mining and Knowledge Discovery*, pp. 1–30, 2023.

[128] E. M. Wickwire, "There is no question about it, sleep disorders increase health care costs," *Journal of Clinical Sleep Medicine*, vol. 17, no. 10, pp. 1971–1972, 2021.

[129] S. Nikkonen, P. Somaskandhan, H. Korkalainen, *et al.*, "Multicentre sleep-stage scoring agreement in the sleep revolution project," *J. Sleep Res.*, vol. 33, no. 1, e13956, Feb. 2024.

[130] S. Redline, R. Budhiraja, V. Kapur, *et al.*, "The scoring of respiratory events in sleep: Reliability and validity," *Journal of Clinical Sleep Medicine*, vol. 03, no. 02, pp. 169–200, 2007. DOI: 10.5664/jcsm.26818. eprint: https://jcsm.aasm.org/doi/pdf/10.5664/jcsm.26818. [Online]. Available: https://jcsm.aasm.org/doi/abs/10.5664/jcsm.26818.

[131] D. Dikeos and G. Georgantopoulos, "Medical comorbidity of sleep disorders," *Curr. Opin. Psychiatry*, vol. 24, no. 4, pp. 346–354, Jul. 2011.

[132] C. A. Goldstein, R. B. Berry, D. T. Kent, *et al.*, "Artificial intelligence in sleep medicine: Background and implications for clinicians," *Journal of Clinical Sleep Medicine*, vol. 16, no. 4, pp. 609–618, 2020.

[133] G. Giray, "A software engineering perspective on engineering machine learning systems: State of the art and challenges," *J. Syst. Softw.*, vol. 180, p. 111 031, Oct. 2021.

[134] H. L. Brennan and S. D. Kirby, "Barriers of artificial intelligence implementation in the diagnosis of obstructive sleep apnea," *Journal of Otolaryngology-Head & Neck Surgery*, vol. 51, no. 1, pp. 1–9, 2022.

[135] T. Grønsund and M. Aanestad, "Augmenting the algorithm: Emerging human-in-the-loop work configurations," *The Journal of Strategic Information Systems*, vol. 29, no. 2, p. 101 614, Jun. 2020.

[136] A. S. Islind and H. V. Hult, "Balancing overreliance and mistrust in data-driven decision making: A critical view on the role of quantified self in diabetes management," in *8th International Workshop on Socio-Technical Perspective in IS Development (STPIS 2022). https://ceur-ws. org*, vol. 3239, 2022.

[137] D. Lee and S. N. Yoon, "Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges," *International Journal of Environmental Research and Public Health*, vol. 18, no. 1, p. 271, 2021.

[138] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, Mar. 2023, ISSN: 1573-7462. DOI: 10.1007/s10462-022-10246-w. [Online]. Available: https://doi.org/10.1007/s10462-022-10246-w.

[139] B. Settles, "Active learning literature survey," 2009.

[140] J. P. Bakker, M. Ross, A. Cerny, *et al.*, "Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: Hypnodensity based on multiple expert scorers and auto-scoring," *Sleep*, vol. 46, no. 2, zsac154, 2023.

[141] S. Prasad, "Designing for scalability and trustworthiness in mhealth systems," pp. 114–133, 2015.

[142] tiangolo, *Fastapi*, 2023. [Online]. Available: https://fastapi.tiangolo.com/.

[143] "Revolution of sleep diagnostics and personalized health care based on digital diagnostics and therapeutics with health data integration.," Tech. Rep., 2021. DOI: 10.3030/965417. [Online]. Available: https://cordis.europa.eu/project/id/965417.

[144] RabbitMQ Contributors, *RabbitMQ documentation*, Version 3.9.0, RabbitMQ, 2007. [Online]. Available: https://www.rabbitmq.com/.

[145] D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," *Linux journal*, vol. 2014, no. 239, p. 2, 2014.

[146] R. Huttunen, T. Leppänen, B. Duce, *et al.*, "A comparison of signal combinations for deep learning-based simultaneous sleep staging and respiratory event detection," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 5, pp. 1704–1714, 2022.

[147] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, Nov. 1971.

[148] R. S. Rosenberg and S. V. Hout, "The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," *Journal of Clinical Sleep Medicine*, vol. 09, no. 01, pp. 81–87, 2013. DOI: 10.5664/jcsm.2350. eprint: https://jcsm.aasm.org/doi/pdf/10.5664/jcsm.2350. [Online]. Available: https://jcsm.aasm.org/doi/abs/10.5664/jcsm.2350.

[149] H. Phan, K. P. Lorenzen, E. Heremans, *et al.*, "L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[150] J. Krause, V. Gulshan, E. Rahimy, *et al.*, "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.

[151] P. Ren, Y. Xiao, X. Chang, *et al.*, "A survey of deep active learning," *ACM computing surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.

[152] B. Holm, *BreathFinder*, version 1.0.0, Aug. 2020. [Online]. Available: https://github.com/benedikthth/breathfinder.

[153] R. J. Adams, S. L. Appleton, A. W. Taylor, *et al.*, "Sleep health of australian adults in 2016: Results of the 2016 sleep health foundation national survey," en, *Sleep Health*, vol. 3, no. 1, pp. 35–42, Feb. 2017.

[154] L.-N. Zeng, Q.-Q. Zong, Y. Yang, *et al.*, "Gender difference in the prevalence of insomnia: A meta-analysis of observational studies," *Frontiers in Psychiatry*, vol. 11, 2020, ISSN: 1664-0640. DOI: 10.3389/fpsyt.2020.577429. [Online]. Available: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2020.577429.

[155] M. Pitkänen, H. Pitkänen, R. K. Nath, *et al.*, "Temporal and sleep stage-dependent agreement in manual scoring of respiratory events," en, *J. Sleep Res.*, e14391, Nov. 2024.

[156] D. Ferretti, A. S. Islind, K. A. Ólafsdóttir, S. Sigurðardóttir, and E. S. Arnardóttir, "The use and quality of 3 nights self-applied home sleep studies," en, *Sleep Med.*, vol. 100, S305, Dec. 2022.

[157] T. U. Wara, A. H. Fahad, A. S. Das, and M. M. H. Shawon, "A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence," *arXiv [cs.LG]*, May 2024.

[158] P. Moridian, A. Shoeibi, M. Khodatars, *et al.*, "Automatic diagnosis of sleep apnea from biomedical signals using artificial intelligence techniques: Methods, challenges, and future works," en, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 12, no. 6, e1478, Nov. 2022.

[159] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, "A systematic review of detecting sleep apnea using deep learning," *Sensors*, vol. 19, no. 22, 2019, ISSN: 1424-8220. DOI: 10.3390/s19224934. [Online]. Available: https://www.mdpi.com/1424-8220/19/22/4934.

[160] D. V. P?S?, "How can we manage biases in artificial intelligence systems – a systematic literature review," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100 165, 2023, ISSN: 2667-0968. DOI: https://doi.org/10.1016/j.jjimei.2023.100165. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096823000125.

[161] R. H. Sprague, "A framework for the development of decision support systems," *MIS Quarterly*, vol. 4, no. 4, pp. 1–26, 1980, ISSN: 02767783, 21629730. [Online]. Available: http://www.jstor.org/stable/248957 (visited on 07/30/2024).

[162] A. X. Garg, N. K. J. Adhikari, H. McDonald, *et al.*, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review," *JAMA*, vol. 293, no. 10, pp. 1223–1238, Mar. 2005.

[163] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial intelligence and human trust in healthcare: Focus on clinicians," en, *J. Med. Internet Res.*, vol. 22, no. 6, e15154, Jun. 2020.

[164] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," en, *Hum. Factors*, vol. 52, no. 3, pp. 381–410, Jun. 2010.

[165] L. Harbarth, E. Gößwein, D. Bodemer, and L. Schnaubert, "(over)trusting AI recommendations: How system and person variables affect dimensions of complacency," en, *Int. J. Hum. Comput. Interact.*, pp. 1–20, Jan. 2024.

[166] B. Holm, G. Jouan, E. Hardarson, *et al.*, "An optimized framework for processing multicentric polysomnographic data incorporating expert human oversight," en, *Front. Neuroinform.*, vol. 18, p. 1 379 932, May 2024.

[167] B. Holm, *Micronyx.is*, `https://micronyx.is`, Platform used for experimental purposes, 2023.

[168] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, Austin, TX, vol. 445, 2010, pp. 51–56.

[169] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: `10.1038/s41592-019-0686-2`. [Online]. Available: `https://doi.org/10.1038/s41592-019-0686-2`.

[170] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947. DOI: `10.1214/aoms/1177730491`. [Online]. Available: `https://doi.org/10.1214/aoms/1177730491`.

[171] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: `https://www.R-project.org/`.

[172] M. Kay, L. A. Elkin, J. J. Higgins, and J. O. Wobbrock, *ARTool: Aligned rank transform for nonparametric factorial anovas*, note. DOI: `10.5281/zenodo.594511`. [Online]. Available: `https://github.com/mjskay/ARTool`.

[173] A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti, "The importance of respiratory rate monitoring: From healthcare to sport and exercise," *Sensors*, vol. 20, no. 21, 2020, ISSN: 1424-8220. DOI: 10.3390/s20216396. [Online]. Available: https://www.mdpi.com/1424-8220/20/21/6396.

[174] N. Mohammadzadeh, S. Rezayi, and S. Saeedi, "Telemedicine for patient management in remote areas and underserved populations," *Disaster Medicine and Public Health Preparedness*, vol. 17, e167, 2023. DOI: 10.1017/dmp.2022.76.

[175] B. Noah, M. S. Keller, S. Mosadeghi, *et al.*, "Impact of remote patient monitoring on clinical outcomes: An updated meta-analysis of randomized controlled trials," en, *NPJ Digit. Med.*, vol. 1, no. 1, p. 20 172, Jan. 2018.

Appendix A

Appendices

.1 ScoreCraft Study Completion Questionnaire

Participant Information

1. **What is your name?**
   (We will not use your personal identity for any analysis or publication, and it will remain completely confidential.)
   Answer: _____
2. **What is your email?**
   (If this is not prefilled, please use the same email you used to log into MicroNyx.)
   Answer: _____
3. **Where are you working?**
   (Center, City, Country)
   Answer: _____

Study Awareness

1. **I heard about this study from:**
   - The invitation email sent to the Sleep Revolution.
   - The talk held by Dr. Erna Sif Arnardóttir at the Sleep Europe conference.
   - The talk held by Dr. Erna Sif Arnardóttir at the ERS conference.
   - Email sent to the members of the ESST.
   - Other: _____

2. **Are you part of the Sleep Revolution?**
    - Yes
    - No

User Experience

This section of the questionnaire is designed to gain insight into the ease of use of the MicroNyx platform.

1. **The system was easy to use.**
   Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
2. **I felt that I could reliably read, interpret, and score the signals in the interface.**
   Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
3. **The scoring recommendations were easy to see.**
   Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
4. **It was clear which scoring recommendations were from a human, and which were from an AI.**
   Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree
5. **I felt the recommendations were helpful when scoring.**
   Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Additional Feedback

If you have any further comments, please write them here. We are very happy to hear your feedback.
Answer: _____