New Pathways for Unsupervised Machine Learning in Digital Health: Applications and Future Potentials for Sleep

---

PhD Dissertation by
Luka Biedebach

Department of Computer Science | Reykjavik University
2025

New Pathways for Unsupervised Machine Learning in Digital Health: Applications and Future Potentials for Sleep

by
Luka Biedebach

Dissertation submitted to the School of Computer Science
at Reykjavik University in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

30st of April, 2025

Supervisors:	Anna Sigríður Islind, Professor, Reykjavik University Erna Sif Arnardóttir, Associate Professor, Reykjavik University
Examining Board:	María Óskarsdóttir, Assistant Professor, University of South Hampton Samu Kainulainen, Adjunct Professor, University of Eastern Finland Alexander Moltubakk Kempton, Associate Professor, University of Oslo
Examiner:	Aleksandre Asatiani, Associate Professor, University of Gothenburg

New Pathways for Unsupervised Machine Learning in Digital Health: Applications and Future Directions for Sleep

Luka Biedebach

April 30st, 2025

Abstract

Unsupervised machine learning has emerged as a powerful method, offering the ability to detect patterns, and reveal hidden relationships in health-related data without relying on manual labels. This thesis aims to explore the existing applications and future potentials of unsupervised machine learning in digital health. There are vast amounts of unlabeled data in medical research, and there is a growing need for digital solutions to tackle health challenges in a growing and aging population. Digital health plays a key role in the paradigm shift of medicine towards predictive, preventive, personalized, and participatory healthcare. By investigating unsupervised machine learning in sleep research, this thesis aims to derive implications for digital health and contribute to the fields of information systems and data science.

Sleep, as one of the pillars of health, has a substantial contribution to various mechanisms in our body, including the brain, hormonal balance, and the cardiovascular system. Good sleep can help to improve overall physical and mental health, while poor sleep is associated with chronic diseases, decreased cognitive function, and a shortened lifespan. Even though it is common knowledge that sleep is important, it can be difficult to maintain good sleep. There can be physical, psychological and lifestyle factors impacting the quality of sleep. Digital health and machine learning can address this issue from various perspectives, such as the efficient diagnosis and provision of treatment options for people with sleep disorders, the collection of longitudinal sleep data, and the analysis of sleep recordings.

This thesis first maps all existing publications on unsupervised machine learning in sleep research and then conducts four case studies with selected unsupervised machine learning methods on different forms of sleep data. The cases use anomaly detection,

dimensionality reduction, clustering, and association rules with data on respiration and brain during sleep, as well as objective and subjective sleep quality assessment, and engagement with a digital therapeutics application. Each case aims to represent a novel method to show the range and diversity of contributions of unsupervised machine learning for sleep and encourage researchers to tread new pathways for digital health. The main contributions of the thesis are outlining the existing applications of unsupervised machine learning in sleep research, exploring and applying different forms of health data, critically discussing the clinical value of unsupervised learning, and, ultimately, finding new pathways for unsupervised learning in digital health.

Keywords: Unsupervised Machine Learning, Digital Health, Sleep

Acknowledgements

I want to show my gratitude to the people who made an important contribution to this research. The work on my papers was always a collaborative effort, and I am truly thankful for the opportunity to work with and learn from such great co-authors. Furthermore, I feel very lucky to be part of the Sleep Revolution Family. I want to say thanks to everyone in the research group, regardless of whether they supported me with advice, ideas, or simply great company during coffee and lunch breaks. I want to dedicate a special thanks to Sigríður Sigurðardóttir, Krístin Anna Ólafsdóttir, and María Óskarsdóttir, who supported me in the journey.

Most importantly, I want to thank my supervisors, Erna Sif Arnardóttir and Anna Sigríður Islind, without whom this work would not have been possible. Thank you for always believing in me, supporting me, and being role models. Working with you has largely shaped who I am and want to be as a researcher, and I will be forever thankful for that.

I want to thank my friends here in Iceland who make it feel like home, and thank my friends at home who still remain by my side even though I am far away. Special thanks to Eva-Maria Meyer, who came to Iceland and was a strong support in the last, intense weeks of submitting my thesis. Lastly, I want to thank my family, who have always been and always will be my greatest supporters. Thanks to my grandparents, my parents, and my sister Janna Fee Biedebach for standing by my side.

Being part of the Sleep Revolution project, this research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 965417. Additionally, the first project was supported financially by the Icelandic Research Fund 2016-2019, no. 174067, Nordforsk 2018-2021, (NordSleep, no. 90458) and the Landspitali University Hospital Science Fund 2019-2020 (no. 893831). Nox Medical (Reykjavik, Iceland) additionally supported the study by supplying the researchers with the A1 sleep recorders and consumables needed for the pediatric sleep studies in Paper 2. The birth cohort study was funded by the European Commission: (a) under the 6th Framework Program (FOOD-CT-2005-514000) within the collaborative research initiative 'EuroPrevall', and (b) under the 7th Framework Program (FP7-KBBE-2012-6; grant agreement no. 312 147) within the collaborative project 'iFAAM'.

Publications List

Publications Included in the Thesis

Publication I

L. Biedebach, D. Ferreira-Santos, M.-A. Stefanos, A. Lindhagen, G. N. Pires, E. S. Arnardottir, and A. S. Islind, "Unsupervised machine learning in sleep research: A scoping review", *Under review after major revisions at SLEEP, Oxford University Press*, 2025

Publication II

L. Biedebach, M. Óskarsdóttir, E. S. Arnardóttir, S. Sigurdardóttir, M. V. Clausen, S. Þ. Sigurdardóttir, M. Serwatko, and A. S. Islind, "Anomaly detection in sleep: Detecting mouth breathing in children", *Data Mining and Knowledge Discovery*, vol. 38, no. 3, pp. 976–1005, 2024

Publication III

L. Biedebach, M. Óskarsdóttir, E. S. Arnardottir, and A. S. Islind, "Two Sides of the Same Pillow: Unfolding the Relationship between Objective and Subjective Sleep Quality with Unsupervised Learning", *Proceedings of the Annual International Conference on System Sciences*, 2023

Publication IV

L. Biedebach, M. Rusanen, B. Þórðarson, E. Arnardóttir, M. Óskarsdóttir, S. Nikkonen, H. Korkalainen, S. Myllymaa, J. Töyräs, S. Kainulainen, T. Leppänen, and A. Islind, "Towards a deeper understanding of sleep stages through their representation in the latent space of variational autoencoders", *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 3111–3121, 2023

Publication V

L. Biedebach, K. Ý. Friðgeirsdóttir, C. Carpinelli, A. P. Isberg, H. Helgadóttir, E. S. Arnardottir, J. M. Saavedra Garcia, and A. S. Islind, "Deriving association rules from user engagement in a digital therapeutics application for sleep improvement", *Under Review at the Journal of Medical Internet Research*, 2025

Other Publications by Author

C. Arnold, L. Biedebach, A. Küpfer, and M. Neunhoeffer, "The role of hyperparameters in machine learning models and how to tune them", *Political Science Research and Methods*, vol. 12, no. 4, pp. 841–848, 2024

M. Ghorvei, T. Karhu, S. Hietakoste, D. Ferreira-Santos, H. Hrubos-Strøm, A. S. Islind, L. Biedebach, S. Nikkonen, T. Leppänen, and M. Rusanen, "A comparative analysis of unsupervised machine-learning methods in PSG-related phenotyping", *Journal of Sleep Research*, e14349, 2024

Courses, Workshops, Demos, and Extended Abstracts

L. Biedebach, D. Gozal, E. Arnardóttir, S. Sigurðardóttir, and A. Islind, "To breathe, or not to breathe through the mouth: Analysing mouth breathing in a pediatric sleep study", *Sleep Medicine*,

vol. 115, S286–S287, 2024

L. Biedebach, M. Óskarsdóttir, E. S. Arnardóttir, and A. S. Islind, "Objective and subjective sleep quality", *Nordic Sleep Conference*, 2023

List of Tables

List of Figures

Abbreviations

AASM American Academy for Sleep Medicine.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DTx Digital Therapeutics.

ECG Electrocardiography.

EEG Electroencephalography.

EMG Electromyography.

EOG Electrooculography.

GAN Generative Adversarial Network.

HMM Hidden Markov Model.

ICA Independent Component Analysis.

MRI Magnetic Resonance Imaging.

OSA Obstructive Sleep Apnea.

PCA Principal Component Analysis.

PLMD Periodic Limb Movement Disorder.

PSG Polysomnography.

REM Rapid Eye Movement.

RIP Respiratory Inductance Plethysmography.

RLS Restless Legs Syndrome.

RQ Research Question.

SDB Sleep-Disordered Breathing.

VAE Variational Autoencoder.

Contents

Chapter 1

Introduction

Health is a global issue, with rising chronic conditions and limited healthcare resources. Sleep, as one pillar of health, is tightly interwoven in this complex environment of health challenges [10], [11]. Digital health may be the pathway to overcome these issues in the future by expanding healthcare coverage, improving decision-making, and empowering patients to take care of their own health [12]. Digital technologies, ranging from mobile applications and wearable sensors to advanced analysis, are currently facilitating a new era of healthcare [13]. Digital health allows contemporary medicine to be predictive, preventive, personalized, and participatory, which is referred to as *4P Medicine* [14]. This paradigm shift simplifies and automates existing processes in the healthcare ecosystem and, at the same time, introduces new opportunities. Predictive medicine aims to anticipate conditions before their onset or recognize them at an early stage. An important technological development is machine learning as a predictive tool for disease detection and risk assessment [15]. The success of machine learning in health care is driven by new possibilities for data collection, as well as storing, sharing, and processing of health data [16]. Preventive medicine is closely linked to this data revolution as well. Increased velocity, volume, and variety of health data enable advanced monitoring and screening [17]. By changing how healthcare is delivered, remote care can be a preventive measure for promoting health-supporting behavior. Personalized medicine paves the way for precision medicine by tailoring interventions and treatments to the needs of an individual and making data-

driven decisions based on their distinct characteristics. Participatory medicine reshapes the relationship between patients and the healthcare sector, as well as the relationship of patients with their own health. Simplified measurement and data sharing between patients and healthcare professionals allow patients to track and view their own health data or receive treatment remotely [18], [19].

At its core, digital health is a collection of technologies, including machine learning, that can improve the prevention, diagnosis, treatment, monitoring, and management of health-related issues [20]. While the methods are steadily evolving and promising a healthcare revolution, their adoption in clinical practice is slow as it requires the behavioral change of millions of physicians [21]. Roy, Meena, and Lim [22] identified the need for labeled data, data quality, and low reproducibility in real-life settings as barriers for moving supervised machine learning applications into clinical practice. Unsupervised machine learning, with its diverse applications and ability to detect patterns in health-related data without relying on manual labels, has the potential to open up new research directions [23]. This thesis aims to explore how different unsupervised machine learning methods can contribute to different areas of digital health.

With unsupervised learning, research breaks free from the limitations of labeled data and opens the door to new insights that human-defined labels may have obscured. Unsupervised learning holds high potential for health-related data, as manual labels are often scarce or unreliable [24]. Various methods achieved major breakthroughs in recent years without or only partially relying on labeled data, such as alphago zero [25], BERT [26], or GPT [27]. Generative models, for example, have gained widespread attention in the general population for their ability to produce realistic images, music, and videos [28]. Still, it is often overshadowed by the success of supervised models. The tables may be turning now, as LeCun, Bengio, and Hinton [29] suggested that unsupervised learning will eventually surpass supervised methods in importance, as it mirrors the way humans and animals naturally acquire knowledge. Their words, *"we discover the structure of the world by observing it, not by being told the name of every object"* [29][p.7], capture the essence of unsupervised learning: the power of learning directly from data without relying on labels. This way, unsupervised learning is able to identify hidden patterns without

making prior assumptions about the data and reducing human bias [30]. This thesis explores unsupervised learning by demonstrating its adaptability, interpretability, and clinical relevance across different health-related challenges in the context of sleep research.

Sleep is an interesting application domain for digital health, as it impacts and is impacted by various processes in the body. Therefore, sleep-related research offers a variety of data spanning different physiological processes, measurement devices, and population groups [31]. Sleep can be measured on a millisecond basis, as well as in a long-term study spanning over years. Everyone needs to sleep, which makes it an impactful area of digital health, that everyone can understand and relate to. Machine learning has been a driving force in the evolution of sleep research [32]. Years of research on machine learning have provided us with the methods to automate the analysis of sleep data, which can significantly reduce the workload of medical professionals and sleep researchers. Common applications are the classification of sleep stages [33] and the automatic detection of respiratory events [34]. Machine learning has the potential to, on the one hand, reduce manual effort for sleep technologists and, on the other hand, analyze sleep in unprecedented depth and precision [35]. Yet, there are challenges for machine learning with sleep data, such as the scarcity and ambiguity of manual labels [36].

Therefore, is unsupervised learning a less explored but highly relevant avenue within digital health in general and sleep in particular. Current areas of concern are the rising need for digital health solutions, the slow adoption of existing methods in practice, and the limited amount of research on unsupervised machine learning. Sleep is both, tightly connected to general health from a medical perspective, but also benefits from digital solutions and unsupervised learning from a technical perspective. The thesis aims to contribute to the body of literature on unsupervised machine learning in digital health by i.) outlining the state of the art of unsupervised machine learning in sleep research, ii.) exploring and applying different forms of health data, iii.) evaluating the contribution of unsupervised machine learning from a clinical perspective and iv.) identifying new pathways for unsupervised learning in digital health. The following sections will provide a fundamental understanding of unsupervised learning, sleep, and medical data and show which challenges arise at the intersection of

these domains. Ultimately, this section will outline the structure of the thesis and define four research questions.

1.1 The Rise of Unsupervised Learning

Machine learning teaches computers to learn from data [37]. In this context, learning means improving the computer's own performance in a given task based on experience by being confronted with data and creating an internal model. This model of the known data is then used to identify patterns or make predictions about new data [38]. This is a fundamental shift in programming. Hence, instead of providing the computer with rules on how to achieve a task, we teach it to derive its own rules. Even though the theoretical foundations for machine learning reach as far back as 1958 [39], its rapid evolution has been driven by the advancements in computational power and availability of large datasets.

There are different learning approaches within machine learning. The most common ones are supervised learning, unsupervised learning, and reinforcement learning. Firstly, *Supervised Learning* is trained on labeled datasets, where each training example is paired with an output label. This way, a mapping from inputs to outputs based on the known data and labels is learned. These models then apply the mapping to predict the most likely label for new unlabeled data [40]. Secondly, *Unsupervised Learning* is trained on unlabeled data. The primary goal is to uncover hidden structures or patterns in the data without predefined outputs. Thirdly, *Reinforcement Learning* is trained through interactions with an environment. A so-called learning agent receives feedback in the form of rewards or penalties, guiding it toward actions that maximize a reward function over time [41].

There are two other common forms of machine learning that are related to but cannot be directly placed in these three categories: semi-supervised and self-supervised learning. Both methods show characteristics of supervised and unsupervised methods. *Semi-supervised learning* bridges the gap between supervised and unsupervised learning by leveraging both labeled and unlabeled data [42]. *Self-supervised learning* trains creates labels from unlabeled data [43].

Unsupervised machine learning can infer patterns within data without reference to known or labeled outcomes [44]. Summariz-

ing from multiple resources, unsupervised machine learning can be defined as *every machine learning method that does not rely on labeled data* [44], [45]. Unsupervised machine learning includes various methods such as clustering, dimensionality reduction, anomaly detection, and association rule learning [46]. Furthermore, different generative models and Hidden Markov Models can be considered as unsupervised machine learning [47]. The different unsupervised learning methods will be covered in the Related Work.

1.2 Sleep as a Diverse Research Landscape

Sleep research represents a diverse and dynamic landscape that spans multiple application areas. Sleep offers a diverse range of digital health data types for machine learning applications [31]. Sleep is a relatable and relevant research topic since we all do it every single day. At its most fundamental level, it remains an unexplored research domain with the potential to reveal new insights into why we sleep and what happens during sleep, addressing the vital question of sleep's crucial role in sustaining life. Sleep disorders cover a broad spectrum of health issues, from breathing and movement disorders to psychological disturbances [48], with the addition that these conditions often intertwine with other health challenges, such as epilepsy [49] and neurodegenerative disorders [50]. Treatment strategies mirror this complexity, ranging from lifestyle interventions provided in person or digitally, cognitive behavioral therapy to medication and specialized medical devices [51]. Finally, the sleep field holds heterogeneous data types, ranging from detailed clinical measurements to simplified tracking at home. Clinical measurement can reveal detailed physiological information on various processes in the body, while wearables provide less detailed information but can track sleep longitudinally [52]. Collectively, these data types offer high potential for unsupervised machine learning and digital health applications to deepen our understanding and management of sleep health.

1.2.1 The Scope of Sleep Research

Sleep research aims to understand sleep and sleep-related disorders. Sleep is a condition of the body and mind, which is characterized by unconsciousness, inactivity of the nervous system, closed eyes, and relaxed muscles [53]. By falling asleep, the human enters a state in which most body functions work differently than from the waking state [54]. In the following we will give fundamental background on sleep and describe the different areas of sleep research. We will then explain different forms of sleep data and relate them to digital health.

Sleep Architecture

There are two different stages of human sleep: Rapid Eye Movement (REM) Sleep and non-REM sleep. The non-REM sleep is further divided into stages N1, N2, and N3, which are characterized by low brain activity, and gradually increasing depth of sleep. The depth of sleep is measured by the arousal threshold, i.e., how likely a person is to wake up. These stages account for 75% to 80% of sleep [53]. Especially deep sleep (N3) is important for the body since is strongly connected with daytime functioning [55] and strengthening the immune system [56]. The REM stage is characterized by high brain activity, rapid eye movement, and an otherwise paralyzed body. In this stage, most dreams are experienced. During REM sleep, the brain exercises important neural connections that are essential for learning and memory [56]. This stage accounts for 20% to 25% of sleep [53]. After falling asleep, we first pass through different non-REM stages, from light sleep to deep sleep, and after approximately 80-100 minutes, the REM stage is entered for the first time. Sleep stages can be visualized in a hypnogram, a graph that shows the progression through different sleep stages over time. A hypnogram of healthy sleep can be seen in Figure 1.1, where the non-REM stages are colored in grey, and REM is colored in red.

The Circadian Rhythm

The circadian rhythm is the *inner clock* of the body, regulating the sleep-wake cycle [57]. It initiates the release of melatonin, a hormone promoting sleep, and lowers the body temperature during the night. The circadian rhythm is endogenous, as it repeats

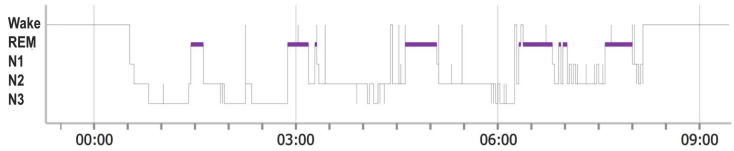


Figure 1.1: A Visualization of the Author's Sleep Stages

in periods of approximately 24 hours, even in constant darkness. However, it can be impacted by environmental factors, such as light, temperature, activity, and nutrition [58]. Being exposed to blue light, e.g., from television or smartphones, in the evening can disturb the circadian rhythm and prevent the release of melatonin, a potential cause why people struggle to fall asleep in the evening [59].

Sleep Hygiene

Taking care of good preconditions for sleep is called maintaining good sleep hygiene. This includes spending as little time in bed as possible, avoiding nicotine, caffeine, and alcohol, and keeping regular sleep schedules [60]. A healthy sleep is characterized by sufficient duration, continuity, and timing [61]. Chronic sleep deprivation or misalignment of sleep to the circadian rhythm, e.g., due to shift work, can have several negative effects on the body, including daytime sleepiness, fatigue, and memory impairment [62]. Sleep deprivation lowers alertness and attention during the daytime [63]. Slower reactions and, in extreme cases, micro-episodes of sleep can lead to accidents e.g., in traffic. A sleep deprivation of 24 hours lowers the reflexes as much as a blood alcohol concentration of 0.5–0.8‰ [64].

Sleep Quality

There is no universally agreed-upon definition of sleep quality or a standard approach to measuring it. It is often characterized by specific aspects of an individual's sleep that can be quantified using parameters such as total sleep duration, sleep onset latency, and sleep efficiency [65]. These metrics are referred to as sleep parameters. Sleep quality can be assessed both subjectively, the subjective experience of sleep, and objectively, the objective mea-

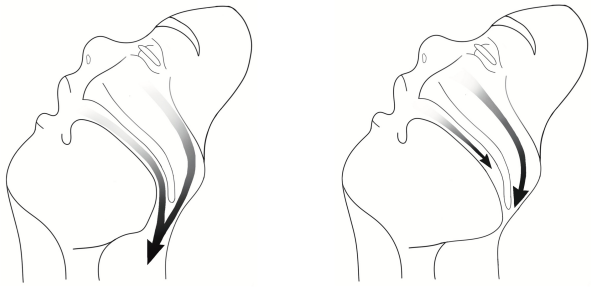
surement of sleep. Information about the subjective experience of sleep is derived from self-reports such as sleep diaries [65]. Subjective sleep quality is not directly linked to any specific sleep parameters but reflects rather a combination of multiple factors [66]. Objective information about sleep is measured, for example, through PSG [67] or wearables [68]. However, often there is no correlation between sleep experiences and measured sleep parameters [69]. Sleep quality is a multidimensional construct, and previous research has shown that subjective and objective sleep quality both impact health [70], [71]. The relationship between subjective and objective sleep quality varies between people of different pathologies [72], medication intake [69], age [73] and gender [74].

Sleep Disorders

The most common sleep disorders are insomnia and obstructive sleep apnea (OSA) [54], which also often coexist [75]. Insomnia is a sleep disorder characterized by difficulty falling asleep, staying asleep, and experiencing non-restorative sleep. Insomnia can stem from a disturbance of the circadian rhythm or an intrinsic sleep mechanism disorder. It can also be caused by sleep disturbance, either by internal factors such as psychological disorders or health conditions or external factors such as noise, temperature, or bed partners. Insomnia can cause fatigue, mood disturbance, daytime sleepiness, and other behavioral problems [76]. Chronic insomnia has wide-ranging impacts on physical and mental health, contributing to conditions such as cardiovascular disease, depression, and cognitive decline. Other sleep disorders are REM sleep behavior disorder [77], Restless Legs Syndrome (RLS) [78], Periodic Limb Movement Disorder (PLMD) [79], narcolepsy, parasomnias, and circadian rhythm disorders [48], [54].

Sleep-disordered breathing (SDB) is a term for all breathing-related sleep disorders ranging from habitual snoring to severe OSA. While SDB often remains undiagnosed, approximately 9% of women and 24% of men in the general population suffer from sleep-disordered breathing [80]. OSA is the most common sleep disorder worldwide. OSA is characterized by temporary interruptions of breathing during sleep [81], caused by an obstruction of the upper airway [82]. It is estimated that globally 425 million individuals have moderate to severe OSA, i.e., with 15 or more

breathing events per hour [83]. A disruption of the airflow for more than 10 seconds is called an apnea. A decrease, but not complete disruption, of the airflow lasting at least 10 seconds is called hypopnea [84]. The severity of OSA is typically indicated by the metric Apnea-Hypopnea Index (AHI), which counts the number of apneas or hypopneas per hour. A relaxation of the muscles in the upper airway causes this obstruction, as can be seen in Figure 1.2.



a.) A Normal Breath b.) An Upper Airway Obstruction

Figure 1.2: A Normal Breath vs. an Upper Airway Obstruction

Apneas lead to an almost complete decline in the airflow, while hypopneas reduce the airflow by 30% or more. Both events occur intermittently and repeatedly during the night, leading to desaturation of the blood oxygen level and hypoxia. This alarms the brain during sleep, causes arousal, and disrupts sleep. For this reason, it is more difficult for people with OSA to get sufficient amounts of restorative sleep. People with OSA usually go back to normal breathing and back to sleep without remembering the disruption, which is why many are not aware of their sleeping disorder. This explains why 80% of the patients affected by OSA remain undiagnosed [85]. Even though short, unnoticeable awakenings like this seem harmless, they can lead to serious health implications. They prevent the patient from entering and staying in REM and deep sleep, which is crucial for restorative sleep. Figure 1.3 shows a hypnogram, i.e., the sleep stages, of an OSA patient derived from the brain activity in a one-night sleep recording. In the hypnogram, it is visible that the person often switches back to wake and enters REM and deep sleep only briefly. OSA

patients only rarely enter deep sleep, and when they do, they are quickly interrupted by arousal. Snoring is often co-occurring or preceding OSA [86].

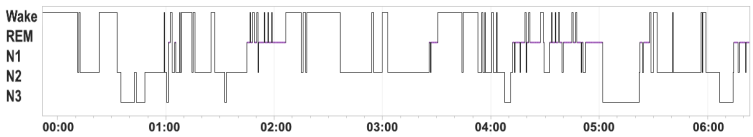


Figure 1.3: The Sleep Stages of an OSA Patient.

If this condition remains undiagnosed and untreated, sleep deprivation and other down-stream effects of breathing disturbance, such as intermittent hypoxia, will seriously affect the health of the patient. OSA can be treated with continuous positive airway pressure [87]. A constant level of pressure is applied to the upper respiratory tract through a nasal or oronasal mask. Contemporary research aims to improve OSA without medical devices, but rather by tackling it at the root and leading patients into to health-supporting behavioral changes, e.g., an exercise-based lifestyle intervention [88], reduce obesity levels [89] or improving muscle tone with myofunctional therapy [90].

1.2.2 Different Forms of Sleep Data

Physiological data can be defined as the collection of measurements that can monitor both the mental and physical status of humans. It captures various physiological processes stemming from various parts of the body. This can be a single measurement, such as e.g., measuring height and weight, a short time measurement, e.g., a reaction test, or a long time measurement, e.g., tracking of blood sugar levels. These examples provide a first impression of the diversity and complexity of physiological data. Sleep data is usually a combination of multiple bio-signals captured during sleep [31].

The most common form of sleep measurement is polysomnography (PSG) [32]. PSG is the continuous recording of physiologic activity during sleep, used to monitor a patient during sleep and diagnose sleep-related conditions. It is a combination of different sensors, measuring different processes of the body. Different forms of sleep studies include different measurement devices. A Type I

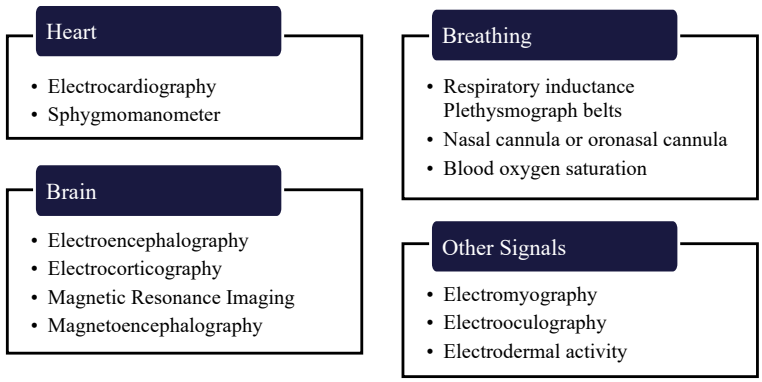


Figure 1.4: Overview of Physiological Data used in Sleep Research [1]

sleep study requires the patient to sleep in a special facility monitored by trained staff. Type II sleep study is typically set up by a medical professional, but the sleep measurement itself can happen at home [67]. Type III sleep study is a simplified form with a reduced number of sensors. Figure 1.5 shows the different sensors included in a Type II sleep study. It includes electroencephalography (EEG) to measure brain activity, electrocardiography (ECG), electrooculography (EOG), and electromyography (EMG), which measure the activity in the brain, the heart, and the muscles [91]. It furthermore includes a nasal cannula and belts measuring the respiration. In the following, each sensor and corresponding data will be described in more detail.

Brain activity during sleep is measured with electrodes which are placed on specific positions of the head. Figure 1.5 shows exemplary positions of electrodes in a PSG recording. The EEG has different characteristics in different sleep stages. The amplitude, frequency, and specific patterns of the EEG signal indicate the current sleep stage [92]. One of these specific patterns, we will refer to as micro features, is the K-Complex which indicates sleep stage N2 and is related to memory consolidation and preventing sleep arousal [93].

The respiratory effort, i.e., the extent of breathing, can be measured with abdominal and thoracic respiratory belts and nasal cannula pressure monitoring [94]. Respiratory Inductance Plethysmography (RIP) belts measure the inflation and deflation of the

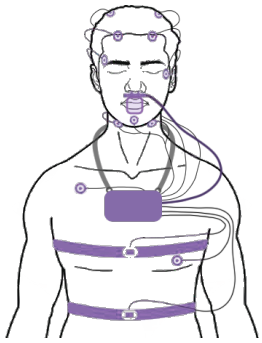


Figure 1.5: Visualization of a PSG Set-up

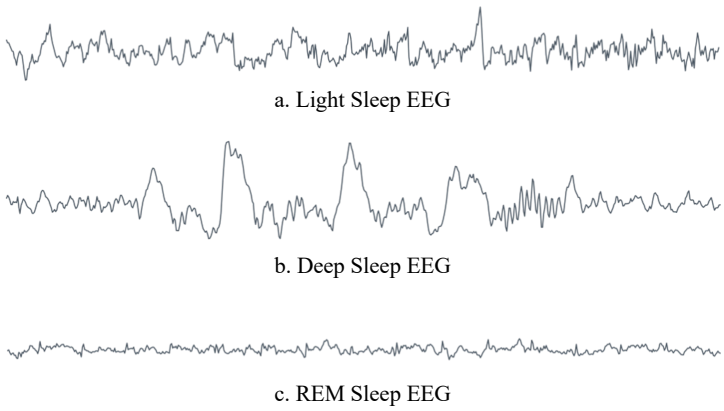


Figure 1.6: Exemplary Sequences of EEG in Light, Deep, and REM Sleep

chest during breathing. One belt is fastened around the thorax, and another belt is fastened around the abdomen, as can be seen in Figure 1.5. The thorax inflates with air in the moment of inhalation and deflates in the moment of exhalation [95]. Hence, the RIP belts indirectly measure the change in lung volume through the strain of the belt [96]. The nasal airflow can be measured with a nasal cannula. It consists of a tube with two openings, which are placed in the nostrils and are fixed with tape. A pressure transducer captures the airflow entering and leaving the nose. Mouth breathing can be captured as well, for example, with an oronasal

cannula or a thermistor. An exemplary sequence of the respiratory signals can be seen in Fig. 1.7, where the thorax movement is colored in light blue, the abdomen movement dark blue, and the nasal flow dark green. The blood oxygen saturation is influenced by breathing, which is why many OSA-related publications use pulse oximeters to detect respiratory events [97]–[100].

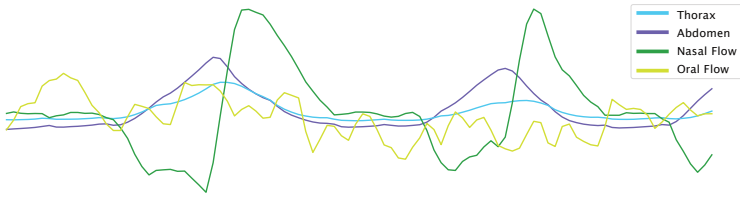


Figure 1.7: An Exemplary 10-second Sequence of Respiratory Signals

Measuring cardiovascular function during sleep can be done with different levels of precision. In PSG, ECG electrodes on the chest offer a precise measurement of heart function during sleep. Blood pressure can also be relevant in research on morning surge [101]. Muscle activity is monitored using EMG, which is typically placed on the chin or on the legs. EMG is useful for detecting conditions like RLS and identifying muscle atonia during REM sleep [102], [103]. Eye movements, which are equally relevant for detecting REM sleep, are tracked through EOG. For this measurement, electrodes are placed close to the eyes. In addition to these, electrodermal activity (EDA) is used to measure skin conductance as an indicator of sweating, offering insights into the autonomic nervous system's activity during sleep [104].

Digital Tools to Measure Sleep

Even though PSG is considered the gold standard for sleep measurement, its complexity and high effort make it impractical for long-term monitoring [67]. In contrast, digital solutions such as wearables or nearables can estimate bed and wake times, nighttime awakenings, and the duration of light and deep sleep based on simplified measurements [68]. They are not as precise as PSG but excel in gathering natural, long-term sleep data, often referred to as 'free-living sleep,' which plays a crucial role in sleep research

[105]. Heart rate and breathing, for example, can be approximated with different wearable devices, including smartwatches and smart rings. They measure heart rate through optical sensors, providing a general but less detailed view of heart activity [106]. Also, nearable devices such as sensors integrated into the bed aim to approximate the heart rate of a person during sleep [107].

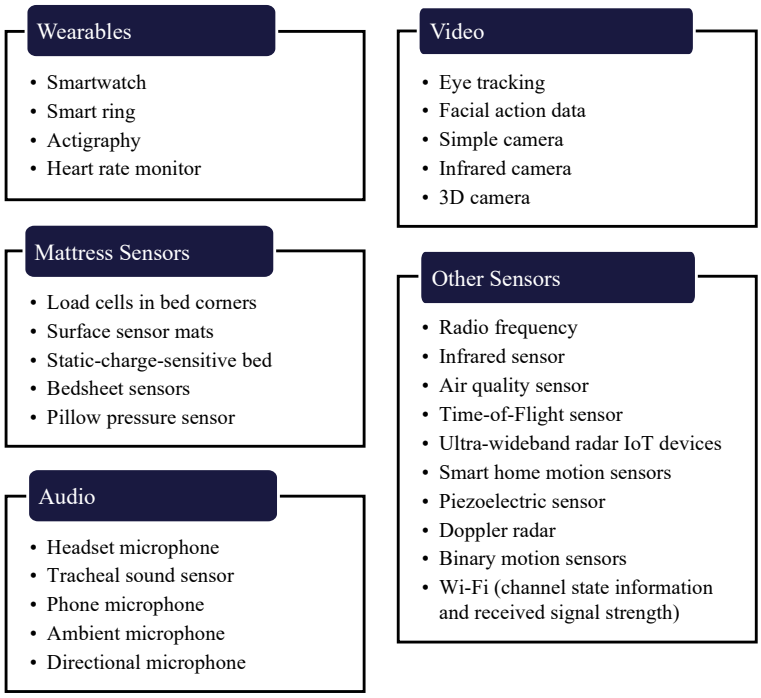


Figure 1.8: Overview of Data from Wearables and Nearables Used in Sleep Research [1]

These measurements can be done with devices that track sleep without being physically attached to the body, such as sleep-tracking mattresses or radar devices. Since these devices do not measure EEG, they estimate sleep and wake times and the sleep stages, e.g., from heart rate, movement, or breathing. They are commonly used for personal health tracking and have proven their value in clinical practice as well [108]. Figure 1.8 shows different forms of tracking sleep and gives examples for specific devices in

each category. There is a new dawn in sleep research, where novel data streams are becoming more interesting and insightful [31].

Smartwatches and rings are used for sleep staging [109]–[111], detecting respiratory events [112] and analyzing sleep patterns [113], [114]. Video has been used to monitor sleep postures [115], movements [116], breathing [117] and heart rate [118] during sleep. Audio measures sleep by placing a microphone on the body or close to the bed. This way, sounds like coughs [119] or snoring [120], [121] during the night can be identified. A different approach is extracting information about sleep from speech during wake, e.g., to predict OSA and other sleep disorders from speech recordings during wake [122]–[124].

An individual's subjective perception of sleep, whether they feel rested or fatigued, can be just as important, or even more important than physiological measurements [70], [71], [125]. One way to measure subjective sleep quality is through Likert-style self-assessments, where participants rate their sleep experience [65]. For a more detailed evaluation, sleep diaries break these assessments down into specific aspects, such as ease of falling asleep or waking up during the night [126]. These evaluations can be recorded digitally in a symptom-tracking mobile application or manually documented in a paper-based sleep diary [127].

Challenges of Sleep Data

There are multiple challenges that rise through the current use of sleep data. These include the high labeling effort, low reliability of labels, and data complexity. Sleep technologists are professionals who can score sleep data manually. They label the signals based on rules defined by the American Academy for Sleep Medicine (AASM) [128]. Possible labels include, e.g., sleep stages, arousal, and breathing events. The scoring requires 2-3 hours of time by a sleep technologist [129]. This limits the number of people with sleep disorders receiving a diagnosis and treatment, leading to long waiting lists for taking a sleep measurement.

The data of PSG recordings can be saved in the .edf standard format, an open-source file format used in the medical sector. It is designed for multi-channel medical time series and allows different sampling frequencies for each signal [130]. Depending on the sampling frequencies and the number of channels used in the PSG, this results in a complex data set. A single signal of 8 hours

with a sampling frequency of 200 Hz results in 5,760,000 values [2]. Hence, one challenge of machine learning in sleep research is data complexity and diversity. To process this large amount of data, we face a trade-off between run-time and the completeness of the data representation. Additionally, the signals have different scales. As some machine learning models are sensitive to different-sized scales, there is the need to prevent the signals with larger scales to out rule the signals with smaller scales by scaling the data to the same range.

Another limitation is that the manual labels of sleep stages are not only time-consuming and expensive but also rely on the subjective judgment of the sleep technologist. The significance of this subjectivity can be observed when comparing the manual labels of two independent sleep technologists when scoring sleep stages. Their inter-rater reliability commonly results in a Cohen's Kappa of 0.71 [36]. Even lower agreement exists between sleep technologists from different laboratories [131] or in populations with sleep disorders [132]. Considering that the manual labeling of sleep data is time-consuming and expensive, and the reliability of these labels is relatively low, it makes sense to explore unsupervised learning that does not train on manual labels.

1.3 Outline and Contribution

This PhD thesis investigates the role of unsupervised machine learning in advancing digital health applications within sleep research. By exploring both existing and potential applications, the work aims to uncover new pathways that can drive innovation in the field of digital health. The thesis aims to contribute to information systems and data science by providing an understanding of the state of the art of unsupervised learning in this area, as well as implementing and evaluating different unsupervised learning methods in different areas of sleep research. Ultimately this thesis aims to, both, connect advancements in unsupervised learning with advancements in digital health, and consider these technical findings from social/human perspective to create a balanced sociotechnical view [133].

We aim to answer the following research questions (RQs):

- **RQ1:** What is the state of the art of unsupervised learning methods in sleep research?
- **RQ2:** Which health data types exist and how can they be integrated in unsupervised learning?
- **RQ3:** How can unsupervised learning make a meaningful clinical contribution?
- **RQ4:** Where can unsupervised learning open up new pathways for digital health?

The following sections will explain the core methods of unsupervised machine learning and review how these methods have been used in sleep research. Then, the methodology of this thesis will be presented. The thesis is built upon a comprehensive literature review and four case studies, which build up on a collection of papers that are all published or under consideration for publication in peer-reviewed outlets. The literature review considers the entire width of methods, data types, and applications of unsupervised machine learning in sleep research. The case studies cover various unsupervised learning methods, including anomaly detection, dimensionality reduction, association rule mining, and clustering, and are examined across multiple areas of sleep research. These areas include breathing during sleep, sleep staging,

objective and subjective sleep quality, and the development of digital therapeutics for sleep improvement. In the methods section, the different data sets and unsupervised learning methods used in the case studies will be described in more detail. Then the results of the review (Paper 1), as well as the results of the case studies (Papers 2, 3, 4, and 5) will be summarized and synthesized. Based on this, the research questions will be discussed, and both theoretical and practical implications for sleep and digital health and for data science and information systems will be derived.

Chapter 2

Related Work

Unsupervised machine learning encompasses a broad family of methods that differ considerably in their learning approaches and applications. Despite this diversity, all these methods share one characteristic: they learn patterns and structures directly from data without the need for labeled outcomes. This characteristics makes unsupervised learning inherently different from supervised learning. While supervised learning has received more attention in existing literature in the past [29], the number of publications is rapidly growing [1].

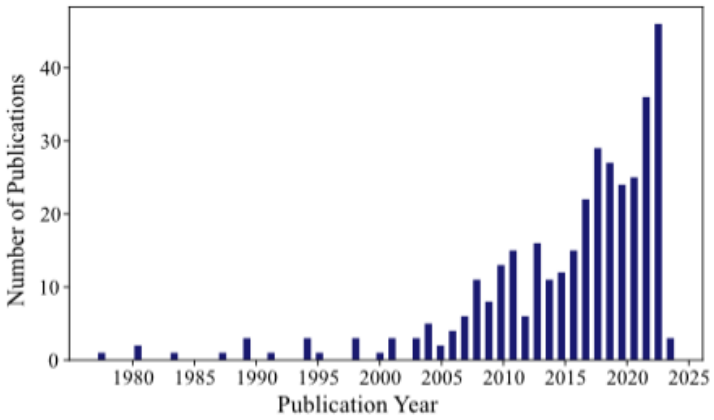


Figure 2.1: Timeline of Publications on Unsupervised Learning in Sleep Research

Figure 2.1 shows the number of publications on unsupervised learning in sleep research from the 1970s until now. The following section will introduce the learning approaches of the most common unsupervised learning approaches, review how they have been applied in sleep research in existing literature, and conclude research gaps.

2.1 Clustering

Clustering is an unsupervised learning method that groups items based on similarity [134]. Items within the same cluster are more similar to each other than they are to items in other clusters. A crucial aspect of clustering, which categorizes it as an unsupervised learning method, is that the cluster labels are not predetermined [135]. Instead, clusters are formed purely based on the intrinsic features of each item. The specific manner in which these features are employed to group similar items varies across different methods [136]. Ghorvei et al. [7] provide a comparative analysis of different clustering methods for PSG-based phenotyping. In the following sections, we will outline the three most common clustering methods.

Partition-based Clustering

In partition-based clustering, the dataset is segmented into distinct groups. The *K-Means* algorithm is a clustering strategy, where the number of clusters, denoted as K, is predetermined [137]. Initially, K centroids are generated at random and represent the centers of these clusters. This establishes an initial clustering configuration, which is later refined through an optimization procedure. During this refinement, the algorithm evaluates the similarity of points within a cluster as well as their distance to other centroids, making adjustments until an optimal arrangement is achieved [138]. Figure 2.2 shows two clusters and the distance of each data point to the centroid. K-Means has been used in sleep research to cluster PSG recordings, where the clusters represented different sleep disorders [139].

Multiple publications use K-Means for the phenotyping of sleep disorders or creating sleep types based on data from wearables, sleep parameters from PSG recordings, or metadata about the

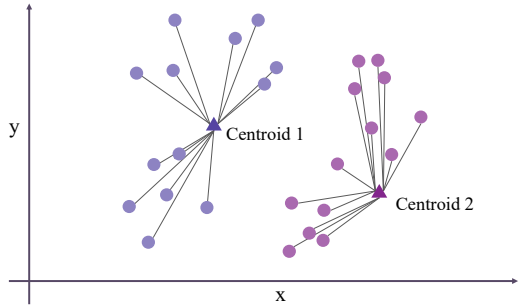


Figure 2.2: Example of Partition-based Clustering

participant [140]–[143]. Park et al. [144] for example, used smartwatch sleep tracking to identify different insomnia types. Other research has used K Means to cluster sleep recordings by their signal quality [145], classify movements [146] and detect snoring [147]. A similar method called, *Fuzzy C-Means*, assigns each data point a probability of belonging to each cluster rather than a definitive assignment [148]. This method has also been widely applied in sleep research, for example, for estimating sleep and wake times based on actigraphy data [149] or extracting features from nearables for classifying sleeping positions [150]. Another application of fuzzy clustering in the scope of sleep-related publications is drowsiness detection. Boyraz, Acar, and Kerr [151] have used fuzzy C-Means clustering to analyze the vigilance of sleep-deprived participants in a driving simulator.

Hierarchical Clustering

Hierarchical clustering builds a hierarchical, tree-like structure [152]. Typically, it is implemented as *Agglomerative Clustering*, where the process begins by considering each element as an independent cluster. Subsequently, the two clusters exhibiting the greatest similarity are identified and combined. This merging continues iteratively until a single cluster encompassing all elements is formed. Figure 2.3 shows how clusters can be layered hierarchically. An alternative strategy is *Divisive Clustering*, where the process begins by considering all elements as one cluster, which is successively divided into smaller clusters. One of the advantages of hierarchical clustering is that it does not require a pre-specified

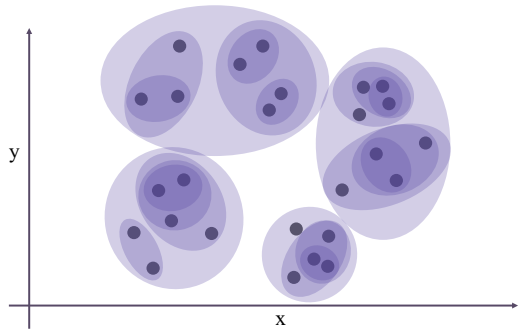


Figure 2.3: Example of Hierarchical Clustering

number of clusters; instead, it allows for the exploration of various levels of grouping. Additionally, the *Dendrogram*, i.e., the resulting tree, offers valuable insights into both the individual elements and the overall hierarchical relationships among them.

There are many publications which have used hierarchical clustering for classifying sleep stages [153]–[156]. In these publications the clustering creates a hierarchical perspective on sleep stages, creating subcategories of sleep stages. Furthermore, hierarchical clustering has been applied to the data from a ballistocardiogram during sleep to model the shape of heartbeats during sleep [157].

Density-based Clustering

Density-based clustering assigns a category to every element based on the density of neighboring elements [158]. In order to assign the categories, we need to define the minimum similarity of two elements to be considered in the neighborhood of each other and the minimum number of elements that have to be in a neighborhood to consider it as densely populated [158]. Then the points will be categorized either as *Core point*, which has more than the minimum number of elements in their neighborhood, as *Border point*, which has less than the minimum number of elements in their neighborhood but at least only one neighboring point, or as *Noise point*, which has no nearby elements. Using these categorizations, we can separate points into clusters. The core points are the core of the cluster, the border points are the borders of the cluster and the noise points do not belong to any cluster. A

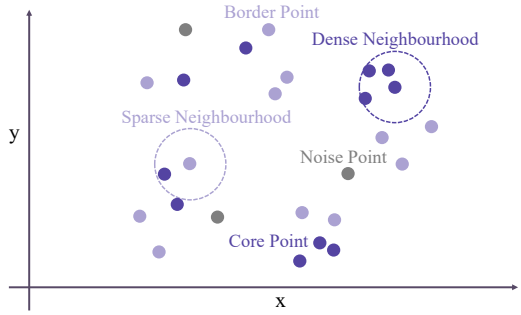


Figure 2.4: Example of Density-based Clustering

visualization of density-based clustering can be seen in Figure 2.4. A well-known implementation of density-based clustering is called *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) [159]. In digital health literature, DBSCAN has been used to identify behavioral risks in older people based on their sleep patterns derived from a mattress sensor [160]. The goal of this publication is, to use DBSCAN as an anomaly detection method by marking the noise points as anomalies. A different application uses density-based clustering to classify sleep stages [161]. Additionally, Yu et al. [162] classify sleep stages with a density-based K-Means clustering algorithm, which follows the K-Means algorithm but uses measures of density and distance to create clusters.

2.2 Dimensionality Reduction

Handling complex data types is a common challenge in sleep research. For instance, PSG records multiple bio-signals simultaneously, typically generating over a hundred measurements per second [2]. Dimensionality reduction aims to simplify such complex datasets while retaining their most essential information [30]. These methods help to enhance data visualization, eliminate noise, and reduce both storage and computational demands. By applying these methods, data visualization becomes more intuitive, noise is filtered out, and storage and computational requirements are minimized. The following sections will introduce Principal Component Analysis (PCA) and Independent Component Analy-

sis (ICA), two widely used methods for dimensionality reduction, alongside Autoencoders, a type of artificial neural network designed to extract compact and meaningful representations of data. Other forms of dimensionality reduction used in sleep research are Singular Value Decomposition, Self-Organizing Maps, and Deep Belief Networks. They are most often used for classification, preprocessing, feature extraction [163], pre-training of a supervised model [164], [165] and data compression [166], [167]. Dhongade and Rao [168] compare different dimensionality reduction methods as a preprocessing step for classifying sleep disorders.

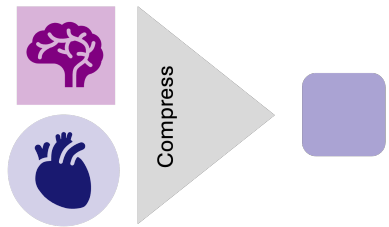


Figure 2.5: Principal Component Analysis

Principal Component Analysis

PCA is a method that can transform data into a set of dimensions known as principal components [169]. Figure 2.5 visualizes the concept of PCA. These components are derived by combining the original features in a way that maximizes the captured variance within the dataset. They are then ranked according to the amount of variance they explain. By selecting only a subset of these components, PCA effectively reduces the dimensionality of the data while preserving its most significant patterns. PCA has been used in sleep research, e.g., to extract features for classifying respiratory events [170]–[172] and sleep stages [173]–[175].

Independent Component Analysis

ICA is designed to identify and separate independent sources within a dataset [176]. Despite the similarity in their names, ICA and PCA serve distinct purposes and rely on different methodologies. Figure 2.6 shows how the goals of ICA are different from

PCA shown in Figure 2.5. While PCA focuses on combining data into components that capture the most variance, ICA instead seeks to disentangle underlying patterns in the data into independent components. A fundamental assumption of ICA is that these sources are statistically independent and exhibit non-Gaussian distributions. ICA is frequently applied to preprocess health data, helping to isolate meaningful physiological processes from unwanted noise. In EEG for example, ICA was widely used in sleep research to separate brain activity from cardiovascular or movement artifacts [177]–[180]. Another common use of ICA in sleep research is decomposing signals to extract breathing or heart rate, for example from a pillow pressure sensor [181] or from video [117], [118]. Lee et al. [182] used ICA to separate the respiration of two people sleeping in one bed, measured with Doppler Radar.

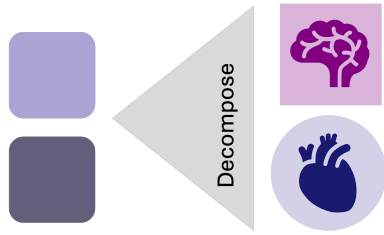


Figure 2.6: Independent Component Analysis

Autoencoder

Autoencoders are a class of neural networks composed of two main components: (i) an encoder and (ii) a decoder [29]. Their architecture is illustrated in Figure 2.7. On the left, the input data, 10-second sequences of a single-channel sleep EEG, enters the network. Given that EEG signals are typically recorded with a sampling frequency of 200 Hz or higher, each sequence consists of at least 2,000 data points [4]. The encoder compresses this high-dimensional input by passing it through multiple layers. The neurons in these layers then learn a compact representation of the data. Once the data reaches the middle of this mirrored architecture, it is stored in a low-dimensional representation. This part of the model is called the latent space. In the example shown in Figure 2.7, each EEG sequence is reduced to just two values.

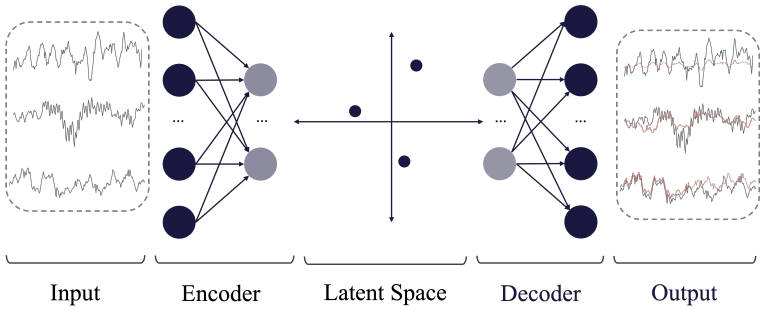


Figure 2.7: Architecture of an Autoencoder

The latent space can be visualized as a coordinate system where each axis represents one extracted feature. While these features are abstract, they may correspond to properties of the signal, such as amplitude and frequency. The decoder then takes this compressed representation and reconstructs the original signal, gradually restoring its structure through successive layers. This can be seen on the right side of the architecture in Figure 2.7. The final output, a reconstructed version of the original input, is compared to the original data, and the difference between them, known as the reconstruction error, serves as an optimization metric for training the network. Through repeated iterations, the neurons adjust their weights to improve reconstruction accuracy, allowing the autoencoder to learn efficient feature representations of the input data.

2.3 Association Rule Mining

Association rule mining is an unsupervised learning method that uncovers hidden relationships among items within a dataset by analyzing how often they appear together [183]. In this context, collections of items that frequently appear together are termed *frequent item sets*. These frequent item sets serve as the foundation for constructing if-then rules. Typically, such a rule is structured so that if a particular item (the antecedent) is present, then another item (the consequent) is likely to be found as well. To derive these rules, three key metrics are evaluated for any combination of items:

Support: This metric determines the frequency with which items co-occur.

$$\text{support}(x_1 \rightarrow x_2) = \frac{N(x_1 \cup x_2)}{N}$$

where $N(x_1 \cup x_2)$ is the number of sets containing both items x_1 and x_2, and N is the total number of sets.

Confidence: This measures the conditional probability that the consequent appears when the antecedent is present.

$$\text{confidence}(x_1 \rightarrow x_2) = \frac{N(x_1 \cup x_2)}{N(x_1)}$$

where $N(x_1)$ is the number of transactions containing x_1.

Lift: This assesses the strength of an association by comparing the observed co-occurrence of items to the probability of their co-occurrence if they were statistically independent.

$$\text{lift}(x_1 \rightarrow x_2) = \frac{\text{confidence}(x_1 \rightarrow x_2)}{\text{support}(x_2)} = \frac{N(x_1 \cup x_2) \cdot N}{N(x_1) \cdot N(x_2)}$$

where $N(x_2)$ is the number of transactions containing x_2.

A widely used method for uncovering these association rules is the *apriori algorithm* [184]. The algorithm leverages the apriori property, which assumes that every subset of a frequent item set must also be frequent [185]. Initially, a minimum support threshold is defined to determine the frequency an item set must reach to be considered frequent. The process begins by computing the support for every two-item combination, retaining only those that meet this threshold. Subsequently, items are iteratively added to these candidate sets, with only those expanded sets that continue to satisfy the minimum support being preserved. This iterative search concludes when no additional frequent item sets can be identified. Association rules are generated by calculating the confidence for all possible directional relationships within each frequent item set. Then, those rules that do not meet the preset confidence level are excluded. Finally, lift is used to distinguish meaningful associations from those arising by chance [183]. There are publications that derive association rules from sleep

data [186]–[191]. All of them were used for explorative analysis of clinical data sets. However, Álvarez, Félix, and Cariñena applied this method to the scoring of respiratory events to identify breathing patterns during sleep [192]. Most of these publications applied association rules on data sets where each person represents one transaction. In contrast, this thesis aims to explore treating every night of a person as one transaction.

2.4 Generative Models

Generative models are designed to learn the underlying patterns in data so they can produce new data that retains the characteristics of the original set. These models can function under both supervised and unsupervised paradigms, with unsupervised methods such as General Adversarial Networks (GANs) being particularly prominent [193]. In a GAN, two components, the *generator* and the *discriminator*, are trained concurrently. The generator's role is to create artificial data by learning from real examples, while the discriminator evaluates both genuine and synthetic data to discern which is which. This adversarial process pushes the generator to produce increasingly realistic data as the discriminator becomes better at detecting subtle differences, leading to a continuous refinement of both components[194].

Generative models have been used in existing research to generate artificial physiological data, which is used to increase the quantity and variety of sleep data used in supervised classifiers. For example, Salazar, Vergara, and Safont use both a GAN and a vector Markov Random Field model to generate artificial ECG data [195]. Also, Latent variable models have been used this way for sleep staging [196] and detecting sleep spindles [197]. Other research has used generative models for data augmentation. Kuo et al. use a GAN to augment sleep data as a preprocessing step for supervised sleep stage classification [198]. Most existing publications aim to generate artificial sleep data as a method to improve the classification performance, e.g., for spindle detection [35], [199], snore detection [120] and sleep staging [200]. Other research aimed to explore the sleep data through these generative models [4] or create art [201]. This thesis aims to contribute to research on generative models by using their embedding space as a way to analyze and explain EEG signals.

2.5 Unsupervised Anomaly Detection

Unsupervised anomaly detection aims to differentiate between what is considered normal and what deviates from the norm. We divide the data into a *normal* and *anomalous* class. In this context, anomalies are data points that diverge significantly from expected patterns. Typically, most of the data falls into the normal category, with only a small fraction being anomalous [202]. The ratio of normal and anomalous points can vary strongly. Anomalies are often characterized by extreme values. However, points within a normal range can also be anomalous. A data point may reside within normal ranges but become unusual when viewed in its context. When looking at sequential data, either individual points can be anomalous, as shown in the example of a respiration signal in Figure 2.8a, or the sequence of multiple points as shown in 2.8b.

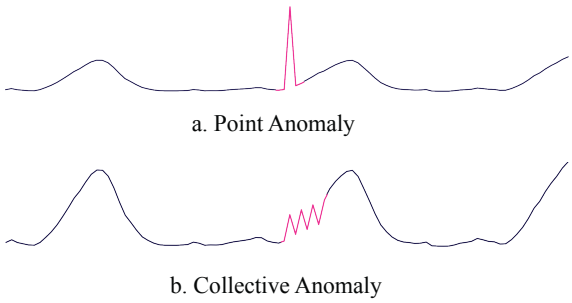


Figure 2.8: Types of Anomalies

The reason why unsupervised machine learning is commonly used for anomaly detection is the high imbalance between normal and anomalous data points. A supervised learning method would learn what normal data and anomalous data look like. However, there are often not enough examples of anomalous data to learn from, and often, the anomalous data points do not necessarily share the same characteristics [203]. As a result, several unsupervised methods are employed—many of which rely on clustering and dimensionality reduction methods introduced earlier [204]. For example, the DBSCAN algorithm can flag points in low-density regions as anomalous, while autoencoders identify anomalies by highlighting data points that incur a high recon-

struction error. Reconstruction-based anomaly detection uses the reconstruction error of autoencoders as an anomaly score. Existing work in sleep research used unsupervised anomaly detection to identify risks based on sleep patterns, such as risk during pregnancy [160] or behavioral risk for the elderly [205], [206]. Other publications used anomaly detection methods for drowsiness detection [207], [208] or detecting anomalous nights [209]. This thesis aims to contribute to research on anomaly detection, by applying reconstruction-based anomaly detection on respiratory data and exploring different ways of training the model, as well as setting the anomaly threshold.

2.6 Other Methods

There are countless other methods relying on unsupervised machine learning. The following section will briefly describe other methods and review related work in sleep and digital health using the methods. A Hidden Markov Model (HMM) is a probabilistic framework that maps an unobservable sequence based on a related, observable sequence [210]. It is based on the assumption that the hidden sequence follows a Markov process, where the likelihood of transitioning to the next state depends solely on the present state. HMM s use transition probabilities to determine the likelihood of moving between states, allowing them to effectively model the temporal structure of sequential data [211]. HMM s have been used in sleep research to model the process of drifting into sleep [212], model sleep transitions [213], and cycling alternating pattern analysis [214].

Unsupervised Domain Adaptation is a method that enables a model trained on a labeled source domain to generalize to an unlabeled target domain. This method has been used in sleep research to apply sleep staging models which have been trained on different public data sets to classify sleep stages on different data sets[215]–[222]. Fan et al. [223] showed that this method improves the classification accuracy when transferring knowledge from big data sets to smaller data sets.

2.7 Research Gaps

Reviewing related work showed how diverse the methods behind unsupervised learning are and how different the applications within this method family can be. Still, most publications on unsupervised learning are either about clustering or dimensionality reduction. Other methods, such as generative models, anomaly detection, or association rules, also show interesting results but have been less often applied in existing work so far. For this reason, there is a need to also consider these less common methods. Even though not all unsupervised learning methods can be covered in the scope of this thesis, the ones implemented as case studies aim to explore different unsupervised methods.

The related work on literature reviews shows that there is no comprehensive review on unsupervised machine learning in sleep. Existing reviews on machine learning in sleep research are focused on one sub-area of sleep research and cover few or no publications on unsupervised learning [224]–[227]. This thesis aims to fill this gap by conducting a scoping review.

Many publications on unsupervised learning in sleep focus on the technical side of the application and do not consider the generalizability of the data set they are using or the clinical meaning of their results [1]. The scoping review on unsupervised learning in sleep research aims to reveal these limitations, and the case studies aim to provide guidance on utilizing sleep data, applying unsupervised methods, and evaluating results in a meaningful way.

Chapter 3

Research Approach

The work toward answering the research questions is comprised of five publications, included in this thesis. These publications answer the provided research question by combining literature research as well as empirical research in the form of case studies. Figure 3.1 shows the literature review as the first phase of the thesis, followed by the four case studies, visualized as the moon as a symbol of sleep.

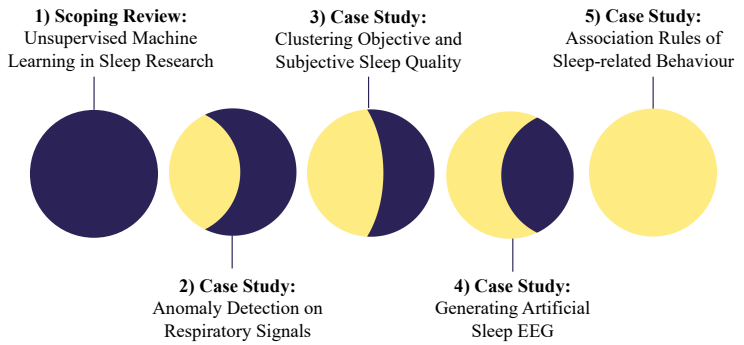


Figure 3.1: The Thesis Divided in Five Phases

The systematic review, here represented as a new moon, reviewed all existing publications on unsupervised machine learning in sleep. This structure of the thesis first provides an understanding of the diverse range of data and methods in this research area

and reveals gaps in existing research. The thesis then dives deeper into specific areas by presenting four case studies [228]. These case studies aim to provide an in-depth understanding of the potentials and limitations of different data types and methods and derive broader conclusions on future directions for unsupervised machine learning in the context of digital health. As proposed by Yin [228], a multiple-cases design can represent a range of variation to provide a more comprehensive perspective on a research question. Therefore, the selected case studies furthermore aimed to represent a broad range of unsupervised methods and data types in order to answer the research questions from diverse perspectives. The topics of all case studies were selected based on ideas for methods and application areas that have not been addressed in the literature before.

The first case study, represented as a growing moon, is focused on unsupervised anomaly detection. The sleep-related application for this case study is sleep-disordered breathing in children. The case study explores different supervised and unsupervised approaches towards time series classification. The paper discusses the need for deep learning and provides guidance on evaluation strategies with health data. The second case study represented as a half moon, analyses objective and subjective sleep quality and personalizes the analysis by clustering the study population based on their perception of their sleep. The paper discusses sleep as a multidimensional construct, as well as the need for analysis on an individual basis instead of one-fits-all approaches. The third case study, here represented as a descending moon, implements a generative model and discusses the use of artificial sleep data and the need for explainable machine learning in clinical decision-making. The last case study, which concludes the thesis and is represented as the full moon, explores sleep improvement in a digital intervention and applies association rules to analyze the participants' engagement with a Digital Therapeutics (DTx) mobile application. In the following, my role as a researcher in each publication will be stated, the different data sets used in each publication will be described and the methodology of the literature review as well as each case study will be explained.

3.1 Researcher's Role

The table below shows how much effort was involved in the various stages of the publication process of the five included publications in this thesis. Starting with the idea, the related literature, data gathering, research design, artifact design, analysis, draft writing, and administration. In the following, each part of the process is shown as defined by the Department of Computer Science at Reykjavik University.

- **Idea**: Crystallizing and formulating a clear and novel research idea alongside research question(s) or hypothesis.
- **Related work and literature**: Reading up on the relevant literature and related work, finding the relevant references as well as putting them together in a coherent manner, alongside building up the research gap.
- **Data gathering**: The gathering of data for the paper.
- **Research design**: Decide on how the data gathering should be conducted (randomized clinical trial, qualitative data gathering, mixed methods, devices used for data gathering or quantitative data gathering, for instance).
- **Artifact design**: In case there is a theoretical model, a method, a digital artifact of some sort (or any software), requirements to be tested, or an algorithm (or machine learning model) that was developed in this category would cover it.
- **Analysis and synthesis**: The analysis of the data alongside the discussion and main contributions are drawn from the analysis.
- **Draft**: The first finished draft of the paper.
- **Administration**: Includes all work with the administration of the publication, such as the submissions of the multiple revisions alongside communication with editors, a major effort in writing the revision comments for the journal papers, and all communication and inclusion of all authors in the various revision rounds.

Paper	Idea	Related work & literature	Data gathering	Research design	Artifact design	Analysis & synthesis	Draft	Admin-istration
Paper [1]	ME	ME	CE	ME	ME	ME	ME	ME
Paper [2]	EE	EE	CE	EE	EE	EE	EE	ME
Paper [3]	ME	ME	CE	ME	ME	ME	ME	ME
Paper [4]	ME	ME	ME	ME	ME	ME	ME	ME
Paper [5]	ME	ME	CE	ME	ME	ME	ME	ME

Table 3.1: Declaration of Authorship Contribution (ME = Main Effort, EE = Equal Effort, CE = Contributing Effort, LE = Learning Effort)

Table 3.1 shows the contribution that I made in the various stages of the publication process for each of the publications. The varying degree of contribution is divided into four categories, including Main effort (ME), Equal Effort (EE), Contributing Effort (CE), and Learning Effort (LE). Each form of author involvement is described in the following, as defined by the Department of Computer Science at Reykjavik University.

- **ME:** Main effort, includes the main effort in the indicated column.
- **EE:** Equal efforts, include that there was a shared equal effort between at least one other author of the paper (this can, for instance, be the case when the work behind the paper was divided or when co-authorship has been equally divided between at least two authors).
- **CE:** Contributing effort, entails important effort, but there is someone else in the author list that delivered the main effort.
- **LE:** Learning effort, includes an effort of a learning character, for instance, by assisting with the data collection or the analysis. At least a LE is needed in all columns to fulfill the Vancouver rules for authorship.

3.2 Literature Research

The literature research is based on a scoping review, a form of systematic review designed to map and analyze a specific research domain [229]. To assess the research landscape in an emerging field like unsupervised machine learning, we performed a qualitative evaluation by outlining its key characteristics and publication trends. The review aimed to capture the entire width of machine learning methods in the entire field of sleep research. This review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines and its extensions for protocols and scoping reviews [230]. The following sections will describe the methodology of conducting the review including the search strategy, inclusion and exclusion criteria, study selection, and data extraction process. The full review

protocol is publicly accessible on the Open Science Framework [231].

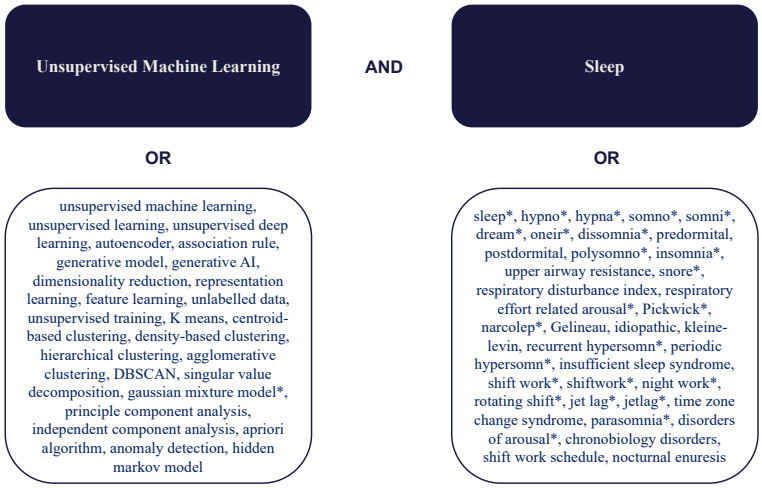


Figure 3.2: Search Terms for Unsupervised Machine Learning and Sleep

3.2.1 Search Strategy

In order to find all existing literature on unsupervised learning in sleep research, a search strategy was composed by combining both components: unsupervised machine learning (methods) and sleep (application domain). No restrictions on the type of sleep study, outcomes, or population were set to increase the sensitivity of the search. The sleep component of the search string was based on the search filter proposed by Pires et al. [232]. An overview of the included search strings and the general search query logic is shown in Figure 3.2. The search query was run in four databases: PubMed, Web of Science, Scopus, and ACM Digital Library. The search strategy was developed for PubMed and then adapted to the other databases' syntax.

3.2.2 Study Selection

The 7043 publications from four different literature databases were de-duplicated. Then, the title and abstract of the remain-

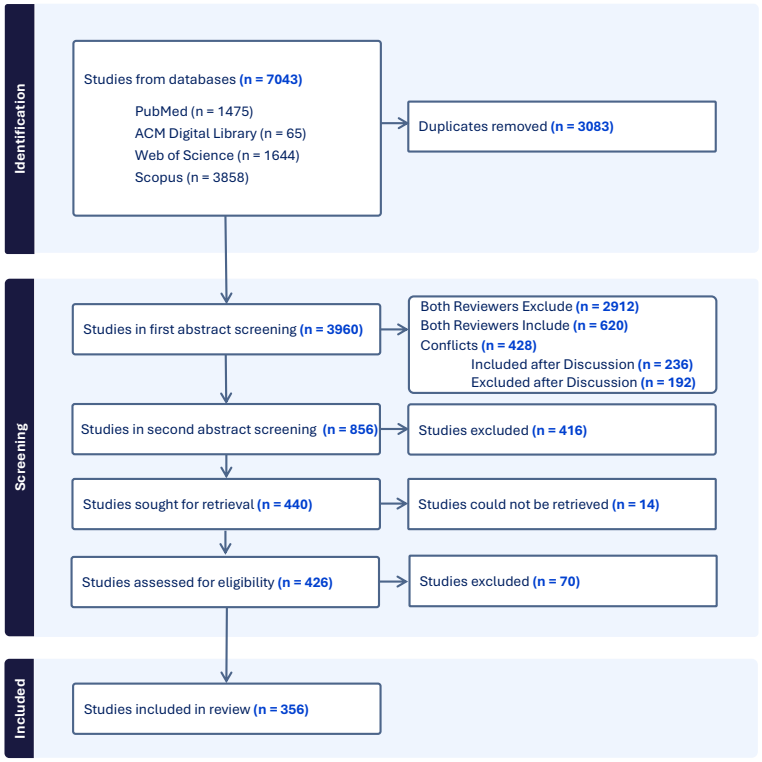


Figure 3.3: Flow Chart According to the PRISMA Guidelines

ing 3960 publications were screened by two independent reviewers. In a second screening round, performed by one reviewer, the publications were further selected on whether they contributed to machine learning. All publications with a contribution only to sleep research were excluded. The full-texts of 426 publications were then reviewed and the data of 356 eligible publications was extracted. A flow chart visualizing this process can be seen in Figure 3.3.

The eligibility analysis for the review was based on several inclusion criteria. Only papers published in English were considered, while publications in other languages were excluded. Additionally, the review focused solely on original research articles, excluding reviews, conceptual or philosophical studies, editorials, letters to the editor, as well as non-peer-reviewed works such as

posters and book chapters. Only studies investigating human populations were eligible, all publications on animal models were excluded. The research had to be primarily related to sleep, defined as having a study population, intervention, explanatory variable, or main outcome associated with sleep. Finally, eligible papers were required to present an application of an unsupervised machine learning method on sleep-related data. This was determined by the abstract mentioning terms such as unsupervised, unlabeled data, or unsupervised training or by identifying specific methods, including clustering, dimensionality reduction, generative models, hidden Markov models, anomaly detection, or association rule learning.

3.2.3 Data Extraction and Analysis

The data extraction was done through a structured, manual process, where each publication was read and the relevant information summarized. Some of the information was extracted as it, while other fields such as *Data Type* or *Sleep Research Area* were categorized. By using both predefined categorical labels with free-text fields, this method enabled both quantitative and qualitative analyses of the reviewed literature. For every paper, metadata—including the first author, publication year, source title, and full reference string—was recorded. Additionally, the country corresponding to the first affiliation of the first author was noted. The review focused on identifying the specific unsupervised machine learning method used, as well as detailing its intended purpose or role in the study. The type of sleep data employed in each paper was documented, along with the practical application of the machine learning method within sleep research. Information regarding the datasets, including population characteristics and sample sizes, was collected as well. An evaluation of the machine learning model's performance was performed by noting the evaluation metrics used at the best-reported performance value. Missing or inapplicable information was consistently recorded as "non-available/not applicable."

3.3 Empirical Research

Case studies are an empirical research method that involves an in-depth, contextual analysis of specific applications within real-

world settings [228]. In the context of this thesis, case studies are employed to implement and evaluate four distinct unsupervised machine learning methods, including anomaly detection, dimensionality reduction, association rule mining, and clustering, on a variety of different sleep data types. This methodology explores how each method performs in practical application, revealing both its strengths and limitations when applied to complex, heterogeneous datasets common in sleep research. By focusing on specific, real-world examples, case studies provide qualitative and quantitative insights that can ultimately help bridge the gap between theoretical advancements and practical implementation.

3.3.1 Overview of Case Studies and Data Sets

In this thesis, I use different forms of sleep data. This includes classical one-night PSG, as well as longitudinal tracking with wearables, digital sleep diaries, and a DTx application. The data stems from four studies with different measurement set-ups and population groups. A timely overview of when these studies were conducted can be seen in Figure 3.4. The characteristics of each data set will be shown in the following. Table 3.2 provides an overview of the different data sets used in this thesis and gives key information on the data collection, population, and measurement channels of each data set.

Figure 3.4: Data Collection of Data Sets Used in this Thesis

The first case study relies on data from a pediatric sleep study with children between the ages of 11 and 14 who are suspected to have sleep-disordered breathing. This study is referred to as *Pediatric Study*. The study was conducted at Landspitali University Hospital, Reykjavik, in 2019. This data set is unique because it has a pediatric focus. Most sleep studies are conducted with adults, even though sleep health is highly relevant for children too. The dataset consists of 111 pediatric sleep recordings, each approximately 8 hours long, collected using PSG. The study focuses on respiratory signals, including oral and nasal airflow, thorax and abdomen movement, oxygen saturation, audio volume,

Data Set	Study Design	Data Collection	Population
Pediatric Study	One night of PSG measurement with an oronasal cannula	Landspitali, University Hospital, Iceland	111 children with suspected sleep-disordered breathing between the age of 11 and 14
50x10 Study	One night of PSG measurement, each scored by 10 different sleep technologists	Akershus University Hospital, Norway and Reykjavik University, Iceland	50 participants of different ages with different comorbidities
Wearable Study	3 months of using a digital sleep diary and smartwatch tracking	Reykjavik University, Iceland	63 participants from the general population
DTx Study	3 months of following a digital intervention and smartwatch tracking	Reykjavik University, Iceland	200 participants with a high body mass index and obstructive sleep apnea

Table 3.2: Data Sets used in this Thesis

and body position. Another unique feature of this study is the oronasal cannula which is not included in a standard PSG setup.

Source	**Variables**
Smartwatch	Sleep hour, wake-up hour, variability, efficiency, regularity, light sleep, deep sleep, duration to wake up, heart rate (avg, min, max), steps, distance, elevation
Sleep Diary	Exercise duration, work day, stress level, nap count, nap duration, drug use, alcohol count, caffeine count
Both	Sleep duration, awake time, awakenings, sleep onset latency

Table 3.3: Variables from the Smartwatch and Sleep Diary

The second case study uses data from longitudinal research with a smartwatch and a digital sleep diary. The participants were asked to use both for 90 days. The study is referred to as *Wearable Study* and was conducted at Reykjavik University from December 2021 until May 2022. The digital sleep diary included a questionnaire about the participants' behavior and well-being during the day filled out in the evening and a questionnaire about the participant's sleep in the morning. This data set is different from the previous case studies as it uses wearable devices instead of clinical measurement devices. It is also different by including tabular data in the form of sleep reporting from the digital sleep diary, which was delivered in a mobile application.

The third case study utilized EEG data from a sleep study with 50 participants. This study is referred to as *50x10 Study* and was conducted in 2021. The participants were selected to have a diverse age range and include participants with different sleep disorders. They collected one night of PSG, which was then been scored by ten different sleep technologists. This unique feature of the data set makes the scoring more reliable by using the consensus of all ten sleep technologists. It is scored for sleep stages and respiratory events. Additionally, it is scored for arousal and micro features within the EEG.

The fourth case study uses data from a study on sleep improvement through lifestyle changes. In particular, sleep was measured before, after, and during a 3 month intervention period. The in-

tervention either took place as an exercise program or a digital intervention. The case study analyzes the sleep and user engagement of the sub-group that did the digital intervention. This study is referred to as *Lifestyle Study* and was conducted from September 2022 until May 2023.

Mission	**Description**
Education	Video, audio, and written content are shown to the user. The educational material is tailored to the specific intervention program and this study includes content about good habits to improve sleep, as well as general lifestyle and health education.
Clinic	Involves logging weight, blood pressure, and other physical measurements, as well as reminders for taking supplements.
Mind	The user is asked to log their stress and energy level, as well as their quality of sleep. This category furthermore involves different breathing and meditation exercises.
Move	The user tracks their activity, which can be anything from walking steps to exercise or sports.
Food	Food intake is tracked and divided into veggies, nuts fruit, snacks, and candy. Beverages are tracked and divided into water and soda. This category also includes unhealthy lifestyle choices such as alcohol and nicotine.

Table 3.4: Different Types of Missions in the DTx Application

The data was collected from three main sources. First, the DTx application recorded user interactions, tracking various behavioral interventions such as monitoring food intake, physical activity, and sleep habits. Second, a digital sleep diary captured self-reported sleep quality and other behavioral factors that might influence sleep. Lastly, smartwatch data provided objective sleep measurements, including total sleep duration, efficiency, and the number of awakenings during the night. The unique feature of this data set is the parallel tracking of sleep data and user engagement with the DTx application.

3.3.2 Case Study: Anomaly Detection

This case study centers around the detection of mouth breathing in the sleep of children. Chronic mouth breathing can have negative effects on the physical and mental health of children but is typically not included in a PSG measurement. This case study aimed to detect mouth breathing based on respiratory signals. The challenge of this data set is, on the one hand, the high imbalance between mouth and nose breathing and, on the other hand, the complexity of the data. Since PSG recordings contain high-frequency data, the signals were downsampled to 10 Hz to reduce computational complexity while retaining essential breathing patterns. Additionally, standardization was applied to normalize the range of different signals

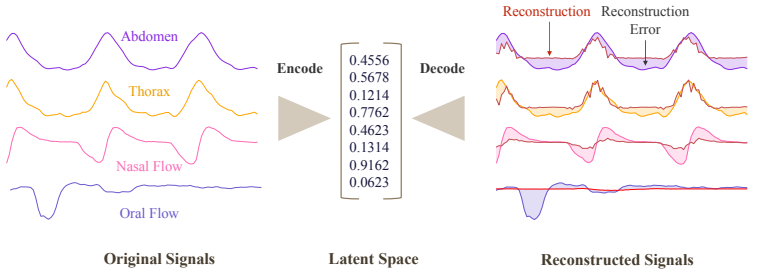


Figure 3.5: Visualization of the Convolutional Autoencoder with Respiratory Signals

This case study compared supervised time series classifiers, such as Time Series Forest, K-Nearest Neighbors with Dynamic Time Warping, and the Multiple Representation Sequence Learner. These models were compared to supervised deep learning models, including a Convolutional Neural Network and a Recurrent Neural Network. Additionally, a simple baseline was tested, extracting features from the time series as for classical machine learning models. A special focus was on reconstruction-based anomaly detection, which aimed to identify mouth breathing without using any labels during the training by taking advantage of this imbalance. For this, a convolutional autoencoder was trained on imbalanced data. It mainly learned the properties of the majority class. It encoded the recorded PSG signals through multiple nonlinear transformations into a low dimensional latent space and

decodes them back in the same way. The reconstructed output, in particular of unknown or anomalous sequences, deviated from the original input. For this reason, the resulting reconstruction error was used as an anomaly score.

To assess the performance of these approaches, the study evaluated models based on precision, recall, and F1-score, with a focus on accurately identifying mouth-breathing instances. To ensure robust generalization, the study employed a leave-one-out cross-validation strategy, training the model on all participants except one and testing it on the excluded individual. This evaluation method prevented data leakage and allowed an assessment of performance across varied individual characteristics.

3.3.3 Case Study: Clustering

In this case study, the sleep quality was captured for 90 consecutive days, both with (i) a smartwatch and (ii) a sleep diary in a mobile application. The data set included 45 participants, who wore the smartwatch on average for 75 days and filled out the sleep diary on average for 40 days. This resulted in 2259 nights in total.

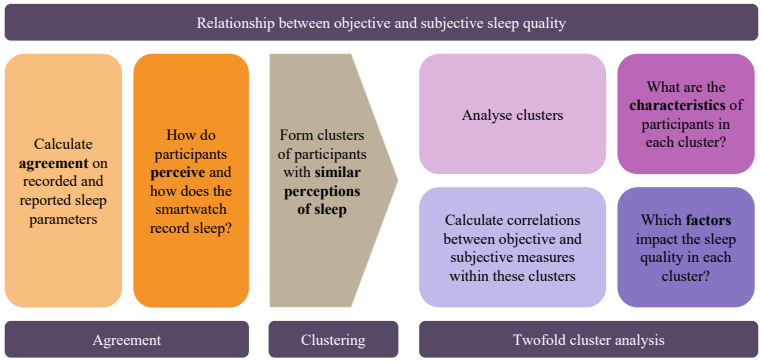


Figure 3.6: Methodology of the Case Study

K-Means clustering divided the data into four clusters. Each cluster calculated Spearman's rank correlation coefficient between possible impacting factors and subjective sleep quality. The rho values represented the positive or negative correlation with the subjective sleep quality, and the p-values the significance of each

correlation. This part of the analysis showed varying relationships between objective and subjective sleep quality within the clusters. Additionally, an analysis of variance (ANOVA) showed which features of the participants within the clusters are significant. This analysis showed varying characteristics of the participants within the clusters. An overview of the method can be seen in Figure 3.6

3.3.4 Case Study: Generative Model

EEG is the main indicator for identifying sleep stages during the night. Both manual scorers and supervised machine learning models mainly rely on EEG, which is why it is one of the most important signals used in sleep research. However, it has not been approached with a variational autoencoder in previous research.

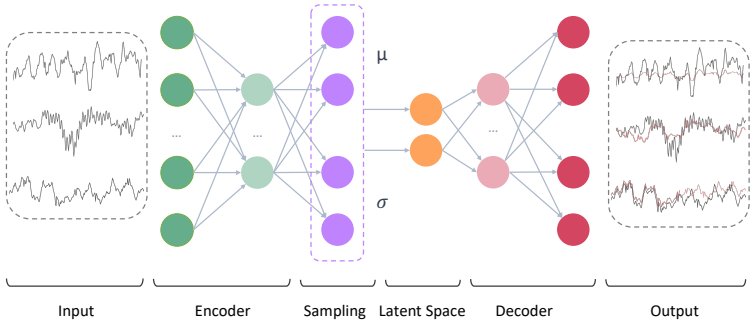


Figure 3.7: Architecture of a Variational Autoencoder

Variational Autoencoders (VAEs) are unsupervised deep neural networks, which are characterized by a continuous latent space. This continuity allows them to act as a generative model. In the context of sleep research, they can be applied to sleep EEG. The F4-M1 channel of 50 sleep recordings (381.13 hours) of both healthy and sleep-disordered participants was used to train a VAE. To lower the data complexity, the signal was downsampled to 64 Hz. High-pass filters with a cut-off frequency of 0.3 Hz and a MinMaxScaling to normalize the data in a range between 0 and 1 were applied. Finally, the data was divided into 10-second segments. With a sampling frequency of 64 Hz this resulted in 143,175 segments of length 640.

The applications of this method were twofold: i.) generate artificial EEG e.g., for creating anonymized sleep data sets, and ii.) get a deeper understanding of sleep stages through their representation in the latent space. The architecture of a variational autoencoder consists of an encoder, the latent space, and a decoder, as can be seen in Figure 3.7. The uniqueness of the variational autoencoder lies within a sampling layer that enforces the continuity of the latent space. VAEs additionally introduce a regularization term in the loss function of the model. The convolutional layer of the autoencoder had 256 filters with kernels of size 5 for the convolutions. It then reduced the length of the sequence with max-pooling. In the following layer, the sequence was flattened into one dimension. In the decoder, a dense layer mirrors the transformation of the sampling layer, and a reshape layer mirrors the transformation of the flattening layer. Then, an up-sampling layer was used to mirror the max-pooling. Finally, a deconvolutional layer with one filter and kernel size 5 brought the data back into their original shape.

Both in the encoder and in the decoder, the activation functions were Rectified Linear Units (ReLU). For the optimization of this model, the weight of the KL divergence was gradually increased from $\beta = 0.01$ to $\beta = 1$ in 100 epochs. For the evaluation, the principles of the Turing Test were applied by confronting a sleep technologist with both real and artificially created EEG sequences. This showed whether the sequences generated by the VAE were realistic. In order to verify that the model could not only generate realistic EEG sequences but also create a meaningful latent space, the characteristics of the generated sequences were manually reviewed with the sleep technologist.

3.3.5 Case Study: Association Rules

This case study explored the relationship between user engagement in a DTx lifestyle intervention program and sleep improvement. To analyze this connection, the study applied association rule mining, specifically the Apriori algorithm, to identify patterns in user behavior, engagement, and sleep quality.

As a first step, the health data from different sources was merged, including the smartwatch, the digital diary, and the DTx application. Then, explorative data analysis was used to get an initial understanding of the user behavior over time and the ad-

herence to the intervention program. The population was further split into two groups based on whether their OSA severity improved over the 12-week intervention period. Since association rules work with categorical data, all numerical columns had to be binned. The binning was manually decided depending on the distribution of each feature. Table 3.5 shows how the sleep parameters were binned.

Feature	**Bins↑**	**Bins↓**
Sleep duration	>8 h	<6 h
Arousal	>5 Awakenings	0 Awakenings
SOL	>1 h	<10 min
Awake	>1 h	-
Screen Time	>8 h	<2 h
Sleep Quality	≥4	≤1
Stay Awake in Bed	>60 min	<15 min
Activity	>10k steps	<1k steps
Stress	≥4	≤1

Table 3.5: Binning of Numerical Features

The data set was transformed into a transaction-based format, where each transaction referred to one night of sleep by one individual. The items in the transaction were the categorical features derived from the smartwatch and the DTx application. In the following an exemplary transaction of Participant n on day m with low sleep duration, a completed education mission ($M_{Education}$) and many awakenings is shown:

X_nY_m={Sleep Duration↓, $M_{Education}$, Awakenings↑}

Adding the days of all participants, the data set resulted in 3200 transactions. Using the Apriori algorithm, frequent item sets were identified. The frequent item sets represent patterns of behaviors that co-occurred regularly among participants. The association rules derived from these patterns indicated how specific interventions, such as mindfulness exercises or dietary tracking, correlated with behavior and sleep parameters. To quantify these relationships, the analysis relied on three key metrics:

- **Support:** how often a specific behavior or intervention was performed alongside sleep improvement

- **Confidence:** the likelihood of sleep improvement following a particular behavior
- **Lift:** the strength of the relationship compared to what would be expected by random chance.

Finally, the association rules by participants with a visible improvement in their sleep quality were compared with association rules by participants with no visible improvement.

Chapter 4

Results

This section presents the results from the literature review and the four case studies, analyzing existing and new approaches to integrating unsupervised machine learning in sleep research and digital health.

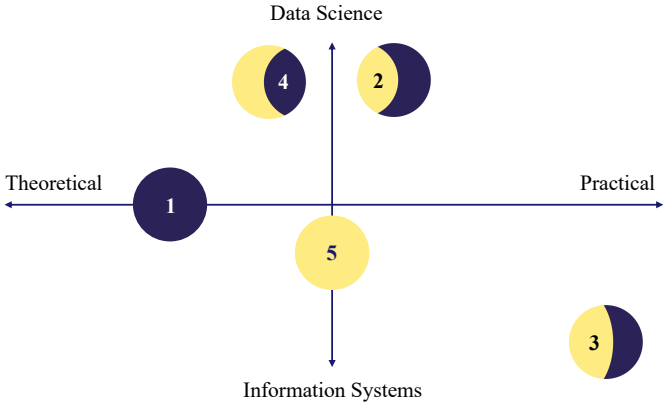


Figure 4.1: Mapping of the Five Publications by their Contribution

By mapping the five publications by their contribution to information systems and data science and the practical or theoretical nature of their contribution, Figure 4.1 shows that this thesis considered the research questions from different perspectives.

4.1 Publication I: Mapping Sleep Literature

Full reference (under review after major revisions):
L. Biedebach, D. Ferreira-Santos, M.-A. Stefanos, A. Lindhagen, G. N. Pires, E. S. Arnardottir, and A. S. Islind, "Unsupervised machine learning in sleep research: A scoping review", *Under review after major revisions at SLEEP, Oxford University Press*, 2025

The scoping review identified 356 papers that use various unsupervised learning methods in sleep research. Analyzing the temporal progression of publications shows that a steep rise in publications on unsupervised learning in sleep research can be observed in the last 10 years, with an even steeper increase in the past 3 years. This indicates that unsupervised machine learning is an emerging method in sleep research. However, the temporal progression also reveals that the foundations for unsupervised learning in sleep research go back to the 1980s.

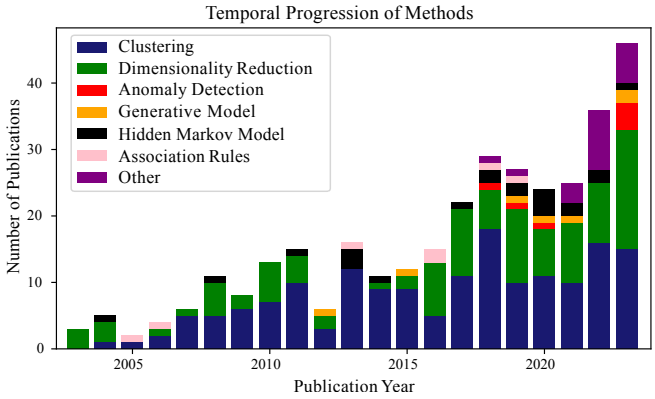


Figure 4.2: Temporal Progression of Publications and Unsupervised Learning Methods

Figure 4.2 shows the number of publications by year and by the unsupervised machine learning method they used. It shows how most publications use clustering and dimensionality reduction, other methods are gaining interest in recent years as well.

Clustering methods were found to be the most commonly used unsupervised machine learning approach, primarily applied to classify sleep stages, detect sleep disorders, and identify sleep phenotypes. Dimensionality reduction methods, such as principal component analysis and autoencoders, have been used for simplifying complex sleep data while retaining essential patterns. Beyond classification, generative models and anomaly detection methods have demonstrated potential in identifying rare sleep events, such as respiratory anomalies, and generating synthetic sleep EEG data to enhance model training.

Figure 4.3: Sankey Diagram Showing the Flow between Data Types and Unsupervised Learning Methods (PSG = Polysomnography,

The scoping review showed that the papers across all sleep applications varied strongly in the way data was collected. The scoping review revealed significant variability in data collection across studies, which were categorized into (i) wearables and nearables, (ii) physiological data, and (iii) metadata and other data. We distinguished between long-term consumer-grade monitoring devices (e.g., wearables, nearables, audio, and video) suitable for home settings and medical-grade devices (e.g., PSG). An additional category included metadata and other unconventional data types. Many studies incorporated multiple data sources, and in such cases, the primary data type was used for the categorization. Figure 4.3 shows the distribution of used data types, as well as the flow between data types and unsupervised methods.

While traditional PSG remains the most common form of sleep assessment, wearable and nearable technologies, combined with machine learning, have expanded the possibilities for long-term, real-world sleep monitoring and are a frequently used data type. Audio was commonly used to identify respiratory events during the night or diagnose OSA based on the speaking voice of a person. Another common data type is video, which was commonly used to detect sleeping positions or classify movements during the night.

The review showed successful applications of unsupervised learning in sleep research but also discovered limitations of existing work. One major limitation of these publications is their generalizability and validity in clinical practice. The review showed that most publications train and validate their models on small populations. Most of them rely on small, homogeneous datasets, often derived from a single population or controlled environments. This way, the methods may show promise in research settings, but only a few studies have successfully translated findings into clinical decision support tools or consumer applications.

4.2 Publication II: Detecting Anomalies in Respiratory Signals

Full reference (published):
L. Biedebach, M. Óskarsdóttir, E. S. Arnardóttir, S. Sigurdardóttir, M. V. Clausen, S. Þ. Sigurdardóttir, M. Serwatko, and A. S. Islind, "Anomaly detection in sleep: Detecting mouth breathing in children", *Data Mining and Knowledge Discovery*, vol. 38, no. 3, pp. 976–1005, 2024

In this case study, we compared different supervised and unsupervised anomaly detection approaches. Breathing through the mouth during sleep has both anomalous properties and clinical relevance, which makes it a suitable application for anomaly detection. Identifying mouth breathing during sleep is currently not included in sleep studies, even though chronic mouth breathing can have negative health implications for children. There are no non-invasive devices to reliably capture mouth breathing.

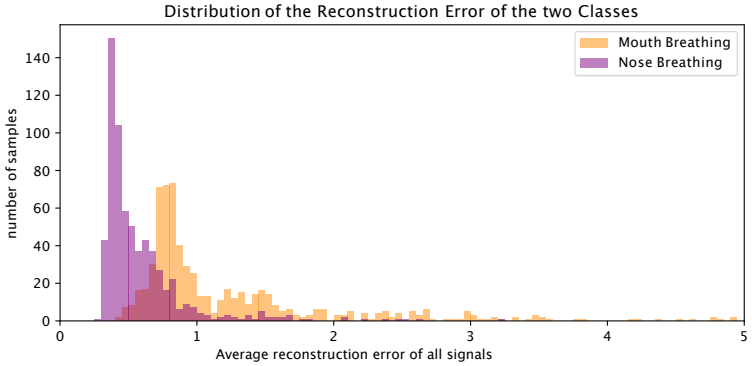


Figure 4.4: Histogram of the Reconstruction Error of the Mouth Breathing and Nose Breathing Sequences

Manual labeling of sleep recordings takes a sleep technologist 2-3 hours per night. The high imbalance towards nose breathing, which is the *normal* airway during the night, makes this classification problem challenging for supervised machine learning models. For this reason, the first proposed application of unsupervised machine learning is the identification of mouth breathing during the

night based on the recording of non-invasive respiratory measurement devices. The goal of this method is to facilitate the diagnosis of chronic mouth breathing and ultimately increase the access to treatment of affected children. Figure 4.4 shows that the reconstruction error of mouth breathing is, on average, twice as high as for the normal airway. Classifying all sequences with a higher reconstruction error than a set threshold as mouth breathing results in a classification accuracy that is comparable to supervised models. This shows, that reconstruction-based anomaly detection is able to identify mouth breathing during sleep without ever learning the properties of this particular class.

The results indicated that reconstruction-based anomaly detection was effective, with an F1 score of 0.508, but it did not outperform supervised learning methods. In particular, feature-based classification with GBM emerged as the most successful method with an F1 score of 0.546, exceeding deep learning models in classification accuracy. This finding challenges the assumption that deep learning always provides superior performance in anomaly detection tasks. Ultimately, the paper provided valuable insights into the comparative strengths of supervised, unsupervised, and reconstruction-based approaches.

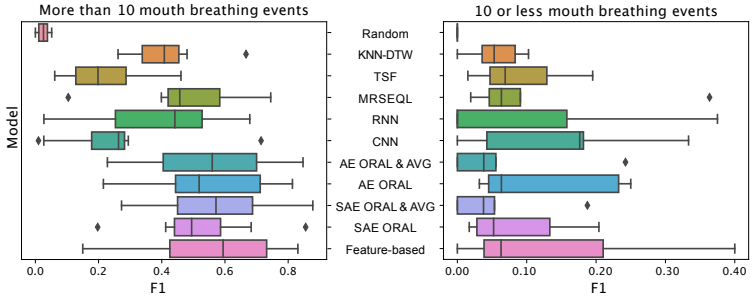


Figure 4.5: Performance on Participants with High and Low Prevalence of Mouth Breathing

The evaluation also showed variability in F1 scores across participants, with some models struggling more with individuals who exhibited fewer instances of mouth breathing. Figure 4.5 shows how the distribution of the classification accuracy varies across individuals. Overall, the results highlight that feature-based classifiers like Gradient Boosting Machine remain superior to deep

learning approaches in this context, challenging the assumption that deep learning always provides the best performance in sleep anomaly detection.

4.3 Publication III: Clustering Objective and Subjective Sleep Quality

Full reference (published):
L. Biedebach, M. Óskarsdóttir, E. S. Arnardottir, and A. S. Islind, "Two Sides of the Same Pillow: Unfolding the Relationship between Objective and Subjective Sleep Quality with Unsupervised Learning", *Proceedings of the Annual International Conference on System Sciences*, 2023

Digital healthcare advancements allow us to take an active part in monitoring and improving our sleep quality. Wearables and digital symptom trackers capture both the objectively measured and subjectively reported sleep quality. The relationship between those two varies between individuals and hence should not be generalized to the whole population.

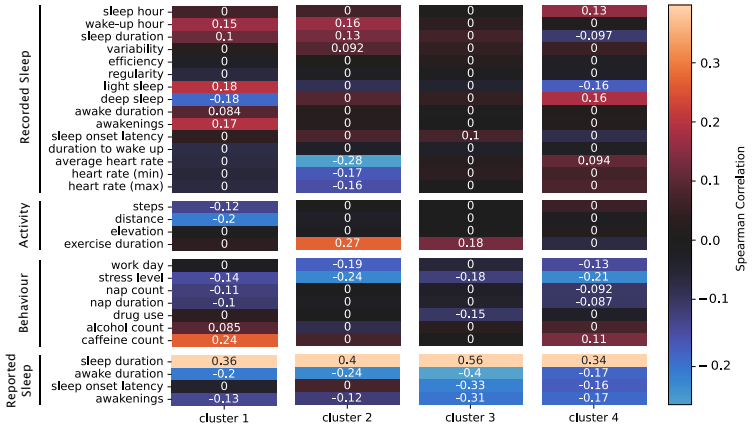


Figure 4.6: Correlations between the Objective Attributes and the Subjective Sleep Quality within each Cluster (non-significant Correlations are Displayed as 0)

Clustering is an area of unsupervised machine learning, which aims to divide data into meaningful groups. The objective of this group of algorithms is usually to increase the similarity between elements within each group. This way, clustering can be used to achieve more individual analysis of heterogeneous data. We

applied K-Means, a partitional clustering algorithm, to identify similar groups of participants within a sleep study. The relationship between objective and subjective sleep quality is highly individual and therefore challenges traditional analysis methods. Clustering participants first on their sleep perception provides a more meaningful analysis of how their objective and subjective sleep quality is correlated. Based on the findings within the clusters, sleep types can be defined which show varying characteristics and correlations to sleep quality.

The proposed method of clustering participants by their perception provided the basis for a more individual analysis of sleep quality. The research showed that analyzing the full population at once did not show any significant correlations between subjective sleep quality and various impact factors. Dividing the population into groups of similar sleep perception allowed us to observe correlations between subjective sleep quality and factors such as exercise, stress level and heart rate. Figure 4.6 shows how four different clusters of participants show different correlations to sleep quality. Clustering the population by their similar sleep perception reveals unique correlations within the clusters. Figure 4.6 shows that only one cluster shows correlations between heart rate-related attributes.

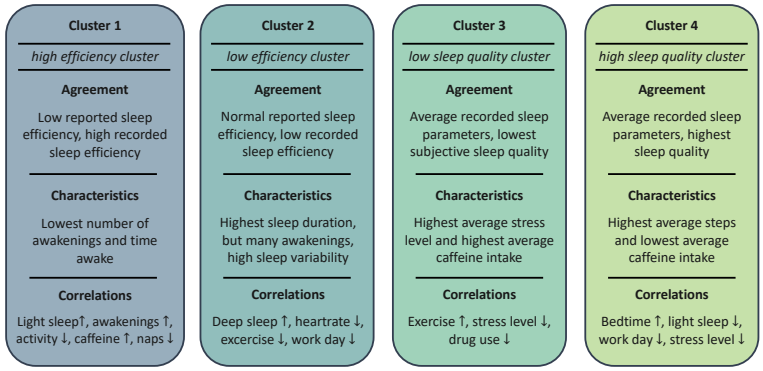


Figure 4.7: Sleep Types Derived from the Clusters

It also shows that one cluster shows almost no correlation between recorded sleep parameters and subjective sleep quality. Additionally, we analyzed the characteristics of the different clusters. This included information about their age, weight, and BMI, as

well as their habits during the day, such as exercise and stress, and their habits regarding sleep, such as regularity or average duration. Combining both results from the correlation analysis and the ANOVA analysis, we defined four sleep types. The first striking difference between clusters was their average sleep efficiency which led to defining a high sleep efficiency and a low sleep efficiency sleep type. The second characteristic that was most visible was the average subjective sleep quality. Clusters three and four had very similar recorded sleep parameters but varied most in their perceived sleep quality. For this reason, we defined them as the high sleep quality and low sleep quality sleep types. All of these sleep types showed different correlations to their subjective sleep quality, and hence different importance to what builds their perception of sleep. This analysis would not have been possible without clustering the participants, which shows that unsupervised learning can be used for more refined and individual sleep analysis.

4.4 Publication IV: Generating Artificial Sleep EEG

Full reference (published):
L. Biedebach, M. Rusanen, B. Þórðarson, E. Arnardóttir, M. Óskarsdóttir, S. Nikkonen, H. Korkalainen, S. Myllymaa, J. Töyräs, S. Kainulainen, T. Leppänen, and A. Islind, "Towards a deeper understanding of sleep stages through their representation in the latent space of variational autoencoders", *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 3111–3121, 2023

The variational autoencoder showed that unsupervised machine learning can be used as a generative method to create artificial EEG data. Figure 4.9 shows a map of artificially generated EEG sequences. The generated sleep EEG sequences were evaluated using a Turing test-style experiment conducted with a sleep technologist. The goal of this experiment was to determine whether the artificial EEG sequences were distinguishable from real EEG recordings. For this, the sleep technologist was first presented with 50 randomly selected EEG sequences, including 4 artificial ones. In the second setting, the sleep technologist was presented with 18 EEG sequences, including 4 artificial ones. The results of the test revealed that the sleep technologist was unable to reliably distinguish between real and artificial EEG sequences in both experimental settings.

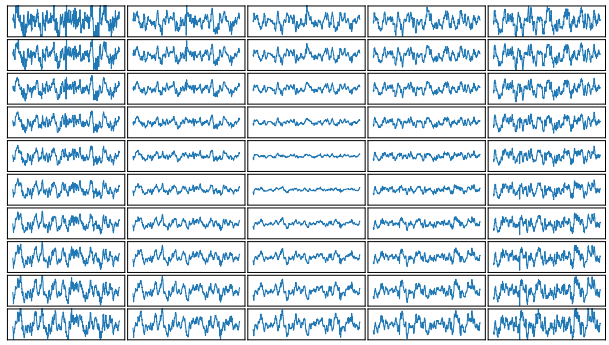


Figure 4.8: Generated Sequences Mapped to their Positions in the two-dimensional Latent Space

Analyzing the different dimensions in the latent space showed that the generated EEG sequences varied systematically depending on their position within the latent space, indicating that the model successfully captured the underlying features of different sleep stages. Certain regions of the latent space corresponded to EEG characteristics typically associated with REM sleep, deep sleep, and wakefulness, suggesting that VAEs can uncover meaningful patterns in sleep EEG data without relying on manually labeled training data. This shows that even a relatively simple neural network architecture can learn key features of sleep EEG patterns in a meaningful way. As the map in Figure 4.9 is only two dimensional, it shows mainly a variation in amplitude and was chosen as a simple visualization. Increasing the dimensionality of the latent space increases the variation within the generated sequences.

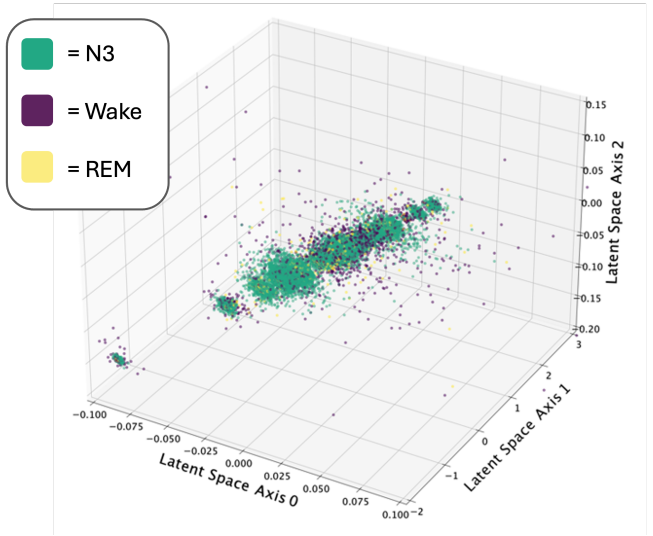


Figure 4.9: Embedding of Real EEG Sequences in a three dimensional Visualization of the Latent Space

The embedding of real EEG sequences in this latent space showed the potential of using the latent space of the variational autoencoder for analyzing sleep stages. Clustering EEG sequences based on their position in the latent space could provide meaningful clusters. Ultimately, this new method could enable an un-

supervised classification of sleep stages. This way, the human bias introduced by the scorers and the human-made scoring rules could be avoided. Additionally, the interpretability of the embedding can provide context to each sequence and could be a new pathway for explainable sleep staging.

4.5 Publication V: Deriving Association Rules from User Engagement

Full reference (submitted):
L. Biedebach, K. Ý. Friðgeirsdóttir, C. Carpinelli, A. P. Isberg, H. Helgadóttir, E. S. Arnardottir, J. M. Saavedra Garcia, and A. S. Islind, "Deriving association rules from user engagement in a digital therapeutics application for sleep improvement", *Under Review at the Journal of Medical Internet Research*, 2025

Analyzing the OSA severity before and after a DTx intervention showed that the sleep did not significantly improve when looking at the entire study population. However, dividing the group into participants who did reduce their OSA severity and participants who did not, provided insights about their differences in behavior.

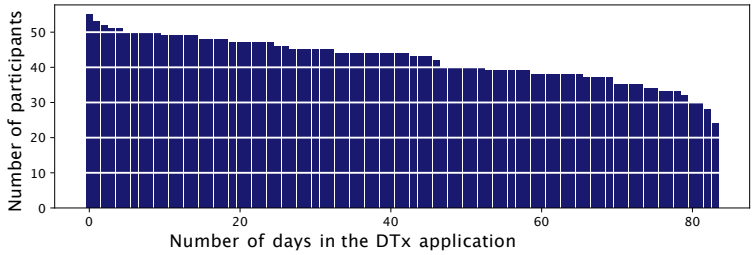


Figure 4.10: Participant retention in the app throughout the 12-week study duration Engagement with the DTx application

It became visible that user engagement decreased over time in both groups, a trend commonly observed in digital health interventions. Despite this decline, 43 out of 55 participants adhered to the study design for more than half of the 12-week period, and 30 participants remained actively engaged for over 80 days. On average, users completed eight missions per day, including activities such as tracking their steps, monitoring food and water intake, and engaging with educational content about sleep. The average engagement with the application overall decreased over time as can be seen in Figure 4.11. Splitting the engagement up by the type of mission the users completed, further differences between

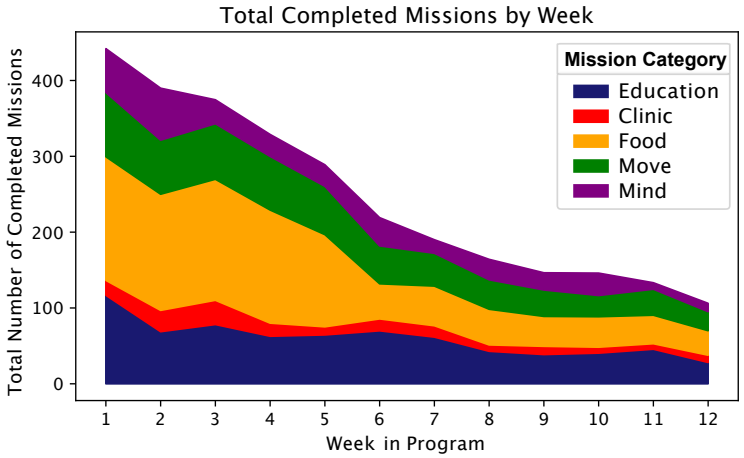


Figure 4.11: Average number of completed missions per week throughout the 12-week program, categorized by mission type

the mission categories arise.

When the population was divided into one group, which showed a reduced OSA severity when comparing the PSG before and after the intervention period, and one group, which did not show improvement, clear differences became visible. The group of participants (n=14) with a reduced OSA severity showed more engagement with the DTx intervention, particularly in the food-related missions, as can be seen in Figure 4.12. The improvement group fulfilled on average 459 missions, while the non-improvement group fulfilled on average 345 missions.

Frequent Item Set	Support	Number of Items
[Awakenings↓, $M_{Education}$, M_{Move}]	0.23	3
[Awake in Bed↓, $M_{Education}$]	0.22	2
[Sleep Duration↑, $M_{Education}$]	0.21	2
[Coach Message, $M_{Education}$]	0.15	2
[Stress↑, M_{Move}]	0.12	2

Table 4.1: Frequent Item Sets

The frequent item sets in Table 4.1 show, for example, that

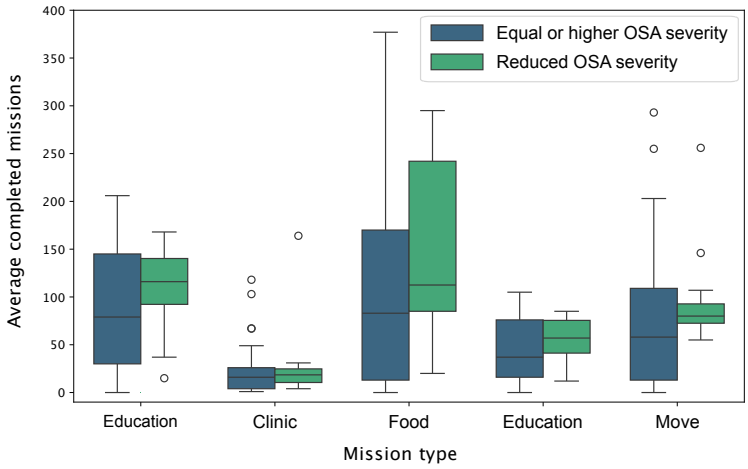


Figure 4.12: Engagement with missions by reduced OSA severity

it is likely that a participant who completed a move mission and stayed only shortly in bed after they woke up is also more likely to complete an education mission and experience fewer awakenings during the night. By applying the apriori algorithm, association rules were identified. The rules revealed several frequent patterns: improved sleep metrics (fewer awakenings, increased sleep duration) often co-occurred with completing education and move missions. Furthermore, participants who completed education and move missions also frequently completed food missions and reported higher sleep quality. Table 4.2 shows examples of the association rules.

Antecedent	Consequent	Sup.	Conf.	Lift
[Awakenings↓, $M_{Education}$]	[Awake in Bed↓, M_{Move}]	0.10	0.40	1.83
[$M_{Education}$, M_{Move}]	[M_{Food}, Sleep Quality↑]	0.12	0.23	1.68
[$M_{Education}$, Screen Time↑]	[M_{Food}, M_{Move}]	0.10	0.79	1.75
[Pain↑, M_{Move}]	[$M_{Education}$]	0.11	0.95	1.55
[Sleep Duration↑, M_{Move}, M_{Food}]	[$M_{Education}$]	0.15	0.94	1.54
[Stress↑, M_{Move}]	[$M_{Education}$]	0.11	0.95	1.54

Table 4.2: Association Rules with Support, Confidence and Lift

Chapter 5

Discussion

By synthesizing the findings from the comprehensive literature review, alongside four case studies, this thesis provides an overview of the landscape of sleep research, highlighting the role of unsupervised learning across various domains within the field of digital health. In the following sections, these results will be connected to the research questions of this thesis and discussed from a sociotechnical lens [133]. By addressing identified gaps and the synthesis of the case study results, directions for new pathways of unsupervised learning will be derived and positioned within the framework of 4P medicine [14]. Based on these contributions, both theoretical and practical implications will be derived. Finally, the limitations of this thesis and future work will be discussed.

5.1 Key Findings

Unsupervised Learning

This thesis showed by reviewing the existing literature that unsupervised learning has already made impactful contributions to sleep research and still has promising applications that have not been explored yet [1]. Historically, unsupervised machine learning is building on the shoulders of giants, capitalizing on the building blocks of mathematics from the last century [39], [138], [233], but has only recently started to gain the attention it deserves. Reviewing the types of methods used in these applications in the scoping review (Paper 1) shows that the majority of unsupervised

learning literature is on clustering and dimensionality reduction. However, there are other promising areas of unsupervised learning, which have been less widely applied in sleep research. The results showed that the benefit of unsupervised models cannot be limited to only refraining from manual labels. They also showed that unsupervised learning can in some use cases, surpass the performance of supervised models and fulfill tasks that are out of the scope of supervised learning. The review furthermore showed how unsupervised methods, for example, clustering or GANs, can be used to improve the classification accuracy of supervised methods.

The results of the four case studies show that unsupervised learning has useful applications in different areas of sleep research, which strongly vary in methodology and research aim, and by that, all contribute in different ways. It shows success in detecting mouth breathing, generating artificial sleep data, identifying sleep types, and analyzing digital interventions. There are applications of unsupervised learning, which could just as well be performed by supervised models. For example, the case study on anomaly detection (Paper 2) showed that unsupervised or semi-supervised methods can be applied as an alternative to supervised anomaly detection to detect mouth breathing and achieve similar accuracy without using labels in the training [2]. In this regard, unsupervised machine learning is beneficial as it saves effort to acquire labeled sleep recordings. Additionally, these models cannot be affected by human bias inherent in the manual scoring. This can let the model seize the true potential of unsupervised machine learning, finding abstract patterns purely based on the data.

However, the applications of unsupervised and supervised learning can not always be directly compared because they are often not suitable for the same tasks. Therefore, exploring the potential of unsupervised learning opens up new applications that have not been considered with supervised machine learning before. The existing findings showed successful applications of unsupervised learning and indicate that there are more applications of unsupervised learning in digital health that are yet to be uncovered. Moreover, the data is yet to be fully carved out, as health data is maturing each day, further capitalizing the need for more research on unsupervised learning. The case on generative models showed that we can generate artificial EEG data in an interpretable and explainable way using unsupervised learning. Without any information about sleep stages in the model training, multiple sleep

stages were represented in the latent space of the autoencoder. The case study using clustering (Paper 3) showed how applying unsupervised learning before traditional analysis can reveal insights that were not visible before [3]. The clustering allowed us to divide the study population into groups which was previously only been done by age or gender. These results indicate that unsupervised learning can open up new research possibilities that are more targeted towards individuals. Similar to objective and subjective sleep quality, this method could be applied to more health conditions and could also, on a grander scheme, be applied to other phenomena that can be measured subjectively and objectively. Hereby this finding is not only limited to sleep research or digital health but could provide pathways for data science and information systems in general.

Therefore, *RQ1*, on the state of the art of unsupervised learning methods in sleep, can be answered by considering the broad spectrum of unsupervised machine learning methods and how each of them takes different roles and can support, replace, or extend the work of supervised models in different ways.

Health Data

Both the review and the case studies gave detailed insights into the different data types used with unsupervised learning methods in sleep research. The review succeeded in highlighting the variety of data types, and the case studies described the selected data types and how they can be used for unsupervised learning in closer detail.

Using simplified measurement devices clearly shows drawbacks in accuracy in comparison to the traditional sleep staging based on EEG but opened new pathways for long-term sleep monitoring, as shown by publications identifying longitudinal patterns in sleep data [107], [206], [234]. The review included many publications using unsupervised machine learning methods to monitor sleep with minimal invasive sleep staging, as shown by devices such as wearables, mattress-integrated sensors, video, audio, and various nearable technologies, which collect sufficient data for basic sleep analysis with minimal intrusion. The thesis showed that unsupervised methods were used to extract information from these low-quality data streams [118], [181], [182]. This usage of unsupervised learning can support the broader adoption of simplified mea-

surement devices in both clinical and home settings, which opens up new possibilities for continuous, longitudinal health tracking. These continuous data streams dramatically increase the volume and variety of available health data, which further benefits unsupervised learning.

Having such a wide choice of data sources raises the question of whether it makes sense to combine multiple data sources. Integrating different types of health data in machine learning models can create new perspectives on sleep, as shown in the case of subjective and objective sleep quality measurements [3] and the connection between smartwatch tracking and user engagement with the DTx application [5]. As sleep and sleep health is a multidimensional construct, where various processes in the body are involved, we need to reflect this also in machine learning applications by integrating different data sources. This applies not only to sleep health but aligns with literature on multi modal health data in general [235]. The richness of various forms of health data and the methods of integrating them open up the possibility of creating machine-learning models that approach health from different perspectives.

The amount of collected health data is steadily growing [236]. The emergence of new health data sources, such as wearables, nearable sensors, and smartphone apps, presents also new challenges. Since health data is highly sensitive personal data, it needs to be treated with care. Unlike traditional clinical data, which is collected in controlled environments with verified devices, wearables are often consumer devices. Furthermore, the data is now collected in the person's home and is being shared over unregulated networks. Therefore, data privacy is a big challenge for new ways of data collection [237], [238]. Another possible weakness of new forms of health data is its reliability, which becomes a critical issue when it is used for monitoring or diagnosing health [239]. If the data is not reliable, basing decisions on it may be harmful to the user.

Based on these findings, *RQ2*, on health data types and their integration in unsupervised learning, can be answered by considering different data forms as different perspectives on health and possibly integrating them for a more complete view. Furthermore, new simplified data types offer new insights but also require caution in terms of data privacy and data reliability.

Integration into Research and Clinical Practice

The state-of-the-art unsupervised learning methods, in combination with new data streams, hold a lot of potential from a technical perspective, but their impact on digital health will be determined by how well the intersection of technology, people, and processes is managed. This sociotechnical perspective has shown that innovation adoption is not automatic in healthcare, as it requires careful alignment with human factors and institutional contexts [240]. In the review and in the case studies showed research that creates clinical value in different ways. In the following, the contribution of unsupervised learning to clinical research, healthcare, and individual health will be discussed. [241]

To ensure a meaningful application of unsupervised learning into clinical research, a fundamental understanding of the methods [242] and medical data [243] is needed. This thesis aimed to create a fundamental understanding of unsupervised methods and medical data through a comprehensive review [1]. One way of ensuring that research is valuable in the medical context is working in interdisciplinary teams. An example is the Sleep Revolution project, where collaborators in academia, healthcare, and industry collaborate in sleep research [129]. All four case studies were conducted in collaboration with researchers from different disciplines, including medicine, biology, medical physics, and sport science, which is reflected both in their conceptualization and evaluation. The case study on anomaly detection, for example, was motivated by the need for better mouth breathing detection, particularly for children, which was identified with the in-depth domain knowledge of a sleep researcher [2]. Furthermore, the falsely classified breathing sequences were manually reviewed with a professional who would review this data in a clinical setting. The conclusion was that most of the false positive sequences either included mouth breathing during wake, slight or short mouth breathing, or bad signal quality. This further confirmed the success of the anomaly detection and, at the same time, undermined the reliability of the manual labels. Similarly, the case study on generative models showed the value of these interdisciplinary research teams, as the artificially generated sleep data was validated by a professional with decades of experience in reviewing real sleep data [4]. These results are in line with Loftus et al. [241], who see large potential for clustering in clinical research, but recommend the

clinical expertise of a medical professional to validate the clusters.

Connecting the findings on unsupervised learning in digital health with information systems theory proposed by Davis [244] shows that the applications need to be useful and easy to use in order to be accepted and integrated into existing processes. If clinicians do not clearly see the usefulness of an unsupervised learning application, they are unlikely to incorporate it into clinical practice [245]. Therefore creating a concrete value, such as detecting anomalous patterns for preventive measures or assisting with easier or more precise diagnosis, is important for convincing clinicians of unsupervised machine learning. At the same time, the applications should be well-integrated into the existing processes and not increase the workload of healthcare professionals. Another important factor for the acceptance of new technologies in healthcare is trust [246]. One way of using unsupervised learning to improve the adoption of AI in healthcare is improving the interpretability of machine learning applications [247]. Explainable AI (XAI) can help to create trust in new technologies and make the usage more intuitive for end-users [248]. Especially in health-related machine learning applications, transparency is important, as the decisions made by an application can directly affect the health of a patient. Generative machine learning models are used in the XAI movement to create visualizations [249]. Other unsupervised methods covered in this thesis, such as hierarchical clustering [250] and dimensionality reduction [251], have been used for explainable decision-making in clinical practice as well.

In conclusion, *RQ3*, on the clinical contribution of unsupervised learning, has to be answered differently depending on whether the application contributes to clinical research or healthcare. In summary, unsupervised learning applications should be developed by an interdisciplinary research team, create trust through explainability, and be well-integrated into existing processes.

5.2 New Pathways for Digital Health

Unsupervised learning is already an established and important method in different areas of sleep research, has been used with various different types of health data, and can make contributions with clinical value when applied right. However, by observ-

ing gaps in existing literature and exploring promising research directions, it became clear that unsupervised learning has yet to realize its full potential. The following section will answer the last research question *RQ4* on where new pathways for unsupervised learning in sleep should be directed and how these findings can be generalized to digital health. These pathways are discussed in the following section by how they can contribute to predictive, preventive, personalized, and participatory medicine.

Unsupervised Learning for Predictive Medicine

The thesis showed how unsupervised learning contributed in different ways to predictive models and, based on that, identified gaps in predictive medicine. The most promising identified future pathways are generative models and unsupervised domain adaptation.

Generative machine learning models recently received an increased interest in the general population. Reviewing the literature on generative models showed new pathways of unsupervised learning with health data. Existing publications showed that generative models learn the underlying distribution of data and can create new synthetic examples, which is a powerful capability for both understanding and augmenting sleep datasets. Synthetic data generation can be used to increase the data set, especially when real data is limited or when simulating rare events. Generative models could also help researchers experiment with hypothetical scenarios, for example, modifying certain signal features to see how sleep architecture might change, by generating new data from the latent space, as proposed in the case study on generative models [4]. Furthermore, unsupervised generative methods can simulate patient profiles or health scenarios, contributing to the development of virtual models of patients used to predict outcomes under different treatments. Ultimately, synthetic data can re-create health data, which reflects important characteristics of a patient's data but does not include personal identifying information, which is an important achievement for patient privacy [252]. Since there have been only few publications and the state-of-the-art in these methods is rapidly evolving, this thesis suggests generative models as a promising new pathway for sleep research. The case study using a VAE emphasizes the importance of transparency and interpretability in machine learning for

healthcare. There is a need for digital health applications to be not only accurate but also understandable by both clinicians and end-users.

Even though typically supervised learning is known for superior prediction accuracy and has been applied in predictive medicine [253], unsupervised methods can be used to enhance the prediction accuracy of supervised models. One method that struck out in the analysis is unsupervised domain adaptation, a method that allows us to apply a machine learning model to sleep data we have no or little access to and train the model on a similar data set. Several publications in sleep research used unsupervised domain adaption, to enhance the accuracy of predictive models. This method has shown success with transferring knowledge from Magnetic Resonance Imaging Magnetic Resonance Imaging (MRI) data stemming from different research centers [254]. Due to the inherent heterogeneity of health data, we propose to pursue this research direction further for data stemming from different participants, devices, or conditions.

Unsupervised Learning for Preventive Medicine

Unsupervised machine learning can contribute to preventive healthcare in different ways. The review showed that continuous monitoring can help with early detection of disease. The thesis identified unsupervised anomaly detection as a potential research direction, especially in combination with emerging data types.

The case study on anomaly detection demonstrated how machine learning methods, including both classic and deep learning models, can identify rare but clinically significant events within highly imbalanced datasets [2]. This method has potential for digital health, particularly in early diagnosis and monitoring of chronic conditions where subtle physiological deviations signal future health risks. The ability to detect anomalies in real-time, using wearable and non-invasive monitoring devices, paves the way for more proactive and personalized healthcare. These methods can be applied to respiratory conditions, cardiac irregularities, or behavioral anomalies in mental health. The challenge of dealing with data imbalance, as seen in the anomaly detection study, also underlines the importance of validation strategies that ensure model robustness across different populations. This finding could apply to any application in digital health that deals with

imbalanced data sets. As suggestions for future research, the work on anomaly detection could be extended by including data from remote monitoring tools as proposed in publications in sleep research [160], [205], [207] and combining multiple data sources.

Unsupervised Learning for Personalized Medicine

Clustering as the most common method of unsupervised learning, showed many strengths within sleep research. Based on the clustering case study, this thesis advocates for personalized medicine in the form of assigning people into clusters with similar health backgrounds.

The case study on sleep quality gave insights into sleep types and the perception of the participants' sleep but also delivered important findings for digital health in general [3]. Wearable devices enable individuals to monitor their health metrics in real time and over extended periods. This shift from clinic-based assessments to home-based monitoring empowers individuals to actively participate in managing their health. The ability to track both objective physiological data (e.g., heart rate, sleep stages) and subjective perceptions (e.g., self-reported sleep, pain, happiness, anxiety) allows for a more holistic understanding of well-being. This concept extends beyond sleep research to other areas of digital health, where integrating various data sources can improve chronic disease management, mental health interventions, and general preventive healthcare. Furthermore, the use of unsupervised learning to identify clusters of individuals with similar health patterns suggests that personalized treatment and intervention plans can be developed across different health domains. For instance, clustering methods can help tailor digital health applications to a certain user group, which is a step towards personalization of healthcare. By uncovering hidden groupings in patient populations, unsupervised learning enables a more personalized approach to healthcare where treatments and health advice can be tailored to the specific data-defined profile of each patient rather than a one-size-fits-all model.

Unsupervised Learning for Participatory Medicine

The case study on analyzing user engagement with a digital health intervention showed how unsupervised machine learning can contribute to participatory medicine. Based on the review and this

case study, this thesis suggests pursuing the analysis of user engagement in digital interventions and highlights association rules as one possible method.

Paper 5 showed that user engagement was an important factor in improving OSA severity with a DTx application [5]. The methodology of using association rules for understanding user engagement and behavior can be used in other DTx applications. This method can be applied to various conditions, such as metabolic disorders, cardiovascular health, and mental well-being. These applications may collect different kinds of data but could benefit in the same way from identifying which features in the DTx application contribute most to improved health outcomes.

5.3 Theoretical Implications

This thesis has several theoretical implications for unsupervised machine learning in sleep and in general digital health. The literature review part of the thesis builds a strong foundation for theoretical implications. By conducting multiple case studies, the thesis furthermore implemented novel methods for handling the unique challenges of digital health data, which is characterized by high dimensionality, noise, and heterogeneity. The theoretical implications derived from the results of all five publications will be discussed in the light of both information systems and data science in the following section.

A key contribution to the field of information systems is the methodology for working with data that reflects the physical and physiological status of humans both through objective and subjective assessment [3]. In other areas of digital health, the same phenomena can be captured both objectively and subjectively, such as stress [255]. While it is up for discussion whether subjective or objective health data are better [256], we aimed to combine both for a more complete view of sleep health. This thesis introduced a novel method for using the gap between objective and subjective data to cluster people and create profiles. Ultimately, this method could be used to enhance the personalization of sleep monitoring and health monitoring in general.

The thesis derived multiple theoretical implications for the field of data science, ranging from deeper analysis pathways for health-related data to improved evaluation procedure that takes

the heterogeneity of patients into account. The case study on exploring sleep stages in the latent space of a VAE shows both how knowledge can be derived from this method, as well as how this method can be evaluated, hereby using an evaluation method based on the Turing test [4]. This can serve as a theoretical foundation for other machine learning publications on generative models. Furthermore, the leave-one-out evaluation and comparison of machine learning models in the case study on anomaly detection makes valuable theoretical implications for machine learning with health data, as well as the trade-off between classical and deep learning. Ultimately, the literature review serves as a strong theoretical foundation for data science research in digital health by mapping, comparing, and discussing unsupervised learning methods [1].

Collectively, these contributions deepen our understanding of unsupervised learning and provide theoretical foundations to drive future innovations in digital health applications.

5.4 Practical Implications

The thesis offers practical contributions by showing different ways of applying unsupervised machine learning methods in the context of sleep and digital health. Both the review and the case studies have practical implications for the successful integration of unsupervised learning in digital health based on the findings on integrating unsupervised learning in sleep research. Starting with a health-related issue that should be both relevant from the perspective of a clinician and feasible from the perspective of a computer scientist, this thesis provides practical guidance for different health data types, preprocessing steps, machine learning methods, and evaluation methods.

First, by working with different types of physiological data, it demonstrates how to combine heterogeneous data sources from wearable sensor outputs to clinical grade recordings. This comprehensive approach not only broadens the scope of available information from traditional measurement devices and new digital tools but also facilitates a more nuanced understanding of health in general and sleep in particular. Analyzing multi-dimensional patient data can create a more complete view of one's health. The case on DTx lifestyle intervention, for example, integrated not only

data from PSG measurement with longitudinal sleep tracking but also explored a completely different form of health-related data: the interaction between the patient and the digital intervention. This new data form, reflecting the engagement of the patients' engagement with their own health, revealed interesting insights and showed how digital health creates new data sources [5]. This paper furthermore proposed to transform this combined data set into a transaction-based format to reflect participants, days and features simultaneously.

Second, this thesis provides guidance on the application of unsupervised machine learning methods on real data sets. Providing robust preprocessing steps, the thesis addresses the challenges inherent in physiological data, such as noise, artifacts, and variations in sampling frequency. Detailed descriptions for filtering, normalization, and dimensionality reduction show how the data are transformed into a suitable format, which is crucial for extracting meaningful features and for improving model robustness in real-world settings.

Third, by comparing different models, the work offers insights into the relative strengths and limitations of various unsupervised methods. Whether using traditional clustering methods or modern deep learning approaches, the comparative analysis guides the selection of appropriate models for specific types of sleep data. By adopting a patient-wise evaluation method, the second paper accounts for inter-individual variability [2]. Evaluating models on a per-patient basis helps ensure that the performance metrics reflect real-world differences across diverse populations, ensuring that the results are both interpretable and clinically relevant. A good example of throughout evaluation of unsupervised learning for healthcare is Miotto et al. [257], who evaluated their work on 76000 patient records predicting 78 different conditions. Moreover, the thesis offered insights into the trade-offs between model complexity and generalizability, demonstrating conditions under which classical anomaly detection methods can outperform more complex deep learning architectures.

5.5 Limitations and Future Work

When analyzing the publications that utilize unsupervised machine learning in sleep research, interesting applications became

visible. However, this analysis also revealed significant limitations of these publications. Many publications validated their work in settings that are not applicable to clinical practice. By testing their methods on small homogeneous populations, their work may not be generalizable to the whole population. This may be enough to prove their method and make a contribution to computer science, but has little value from a health perspective. Interdisciplinary collaboration is essential for translating these methods into real insights with clinical and physiological significance. This thesis advocated for bridging the gap between health expertise and machine learning expertise and aims to provide a foundation to facilitate collaboration across disciplines.

A recurring challenge across the studies is the reliability of recorded data. Physiological signals are inherently complex and heterogeneous. They are prone to measurement errors which can result in missing data or distorted data. Another major challenge is the inter-individual variability and the influence of environmental factors, which can result in substantial differences in the data. Moreover, the reliance on outdated or biased datasets, such as those predominantly featuring young and healthy individuals, limits the generalizability of findings. Due to the absence or low reliability of ground truth labels, the validation of unsupervised learning is limited. Since unsupervised methods are often considered when labeled data is scarce or unreliable, the evaluation based on these labels is inherently flawed. In detecting mouth breathing, for example, defining the exact onset and offset of events remains difficult, even for human reviewers, leading to potential inconsistencies.

Generally, when applying unsupervised machine learning methods, it is difficult to validate and compare the performance of a model. Unlike in classification, there is no straightforward metric to quantify performance in some unsupervised models. In this way, unsupervised models are inherently different from supervised models, as their strength is in exploring new patterns, clusters, or associations instead of recognizing predefined classes. Metrics such as reconstruction error, silhouette scores, or measures of cluster cohesion can quantify a model but may not fully capture the practical utility or interpretability of the model.

A limitation of existing work on unsupervised learning in sleep research is the generalizability of findings. In sleep staging particular, most publications used the Sleep-EDF dataset, a standard

benchmark developed in the 1980s that mainly represents healthy, young Caucasian individuals. This narrow focus raises concerns about the dataset's ability to generalize across diverse populations, particularly in clinical contexts where a broad spectrum of ages, genders, and ethnicities is essential for accurate diagnosis. Moreover, the incremental gains in performance achieved on this dataset may not translate into meaningful clinical improvements. Similarly, the validation of sleep staging models largely on healthy subjects further limits their practical utility. By excluding individuals with sleep disorders, who are most in need of effective diagnostic tools, current approaches risk being not applicable in real-world clinical scenarios. Even though the case studies in this thesis aimed to select appropriate data sets and do meaningful validation, the generalizability of the applied methods to different measurement devices, time frames, or populations is not addressed, which poses a limitation of the case studies. Future research must therefore incorporate more diverse populations to ensure that advancements in machine learning yield robust and clinically relevant outcomes. These challenges underscore the necessity for incorporating domain-specific insights and developing robust, explainable models that can effectively handle the variability and complexity inherent in physiological data.

The thesis is limited by the amount of conducted case studies, since not every identified future research area was implemented. However, the remaining ones were stated as potential pathways and can be explored in future research. Another topic, which was outside of the scope of this thesis but stirred interest for future research was the potential of self-supervised and semi-supervised learning.

Chapter 6

Conclusion

This thesis demonstrated the diversity of research on unsupervised learning in sleep research. Together, the five included papers provide a new perspective on digital health. By moving beyond manual labels and predefined categories and embracing data-driven discovery, the unsupervised methods reviewed and implemented in this thesis guide the way to preventive, precision, and patient-centered care in the digital health era. The literature review, Paper 1, showed that there is a rising interest in unsupervised learning and that the trend goes towards diverse unsupervised methods and using diverse data types. By exploring unsupervised methods such as association rules, generative models, or unsupervised domain adaptation, new research directions, which cannot be covered with supervised models are entered. The case studies, Papers 2 to 5, put promising research into practice and provide guidance on the preprocessing, model selection, and evaluation of these different methods with health-related data. The case studies aim to make clinically meaningful contributions using unsupervised methods and show their potential and challenges, which can be used as a basis for further research in this area. The combined results of these papers outlined existing work of unsupervised machine learning in sleep research, explored and applied various health data, reflected on the clinical value of unsupervised learning, and, ultimately, found new pathways for unsupervised learning in digital health.

Bibliography

[1] L. Biedebach, D. Ferreira-Santos, M.-A. Stefanos, *et al.*, "Unsupervised machine learning in sleep research: A scoping review", *Under review after major revisions at SLEEP, Oxford University Press*, 2025.

[2] L. Biedebach, M. Óskarsdóttir, E. S. Arnardóttir, *et al.*, "Anomaly detection in sleep: Detecting mouth breathing in children", *Data Mining and Knowledge Discovery*, vol. 38, no. 3, pp. 976–1005, 2024.

[3] L. Biedebach, M. Óskarsdóttir, E. S. Arnardottir, and A. S. Islind, "Two Sides of the Same Pillow: Unfolding the Relationship between Objective and Subjective Sleep Quality with Unsupervised Learning", *Proceedings of the Annual International Conference on System Sciences*, 2023.

[4] L. Biedebach, M. Rusanen, B. Þórðarson, *et al.*, "Towards a deeper understanding of sleep stages through their representation in the latent space of variational autoencoders", *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 3111–3121, 2023.

[5] L. Biedebach, K. Ý. Friðgeirsdóttir, C. Carpinelli, *et al.*, "Deriving association rules from user engagement in a digital therapeutics application for sleep improvement", *Under Review at the Journal of Medical Internet Research*, 2025.

[6] C. Arnold, L. Biedebach, A. Küpfer, and M. Neunhoeffer, "The role of hyperparameters in machine learning models and how to tune them", *Political Science Research and Methods*, vol. 12, no. 4, pp. 841–848, 2024.

[7] M. Ghorvei, T. Karhu, S. Hietakoste, *et al.*, "A comparative analysis of unsupervised machine-learning methods in PSG-related phenotyping", *Journal of Sleep Research*, e14349, 2024.

[8] L. Biedebach, D. Gozal, E. Arnardóttir, S. Sigurðardóttir, and A. Islind, "To breathe, or not to breathe through the mouth: Analysing mouth breathing in a pediatric sleep study", *Sleep Medicine*, vol. 115, S286–S287, 2024.

[9] L. Biedebach, M. Óskarsdóttir, E. S. Arnardóttir, and A. S. Islind, "Objective and subjective sleep quality", *Nordic Sleep Conference*, 2023.

[10] D. C. Lim, A. Najafi, L. Afifi, *et al.*, "The need to promote sleep health in public health agendas across the globe", *The Lancet Public Health*, vol. 8, no. 10, e820–e826, 2023.

[11] V. K. Chattu, M. D. Manzar, S. Kumary, D. Burman, D. W. Spence, and S. R. Pandi-Perumal, "The global problem of insufficient sleep and its serious public health implications", in *Healthcare*, MDPI, vol. 7, 2018, p. 1.

[12] U. Varshney, "Mobile health: Four emerging themes of research", *Decision Support Systems*, vol. 66, pp. 20–35, 2014.

[13] A. I. Stoumpos, F. Kitsios, and M. A. Talias, "Digital transformation in healthcare: Technology acceptance and its applications", *International journal of environmental research and public health*, vol. 20, no. 4, p. 3407, 2023.

[14] M. Flores, G. Glusman, K. Brogaard, N. D. Price, and L. Hood, "P4 medicine: How systems medicine will transform the healthcare sector and society", *Personalized medicine*, vol. 10, no. 6, pp. 565–576, 2013.

[15] Y.-K. Lin, H. Chen, R. A. Brown, S.-H. Li, and H.-J. Yang, "Healthcare predictive analytics for risk profiling in chronic care", *Mis Quarterly*, vol. 41, no. 2, pp. 473–496, 2017.

[16] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: Management, analysis and future prospects", *Journal of big data*, vol. 6, no. 1, pp. 1–25, 2019.

[17] T. Hulsen, S. S. Jamuar, A. R. Moody, *et al.*, "From big data to precision medicine", *Frontiers in medicine*, vol. 6, p. 34, 2019.

[18] M. J. Ball and J. Lillis, "E-health: Transforming the physician/patient relationship", *International journal of medical informatics*, vol. 61, no. 1, pp. 1–10, 2001.

[19] S. Shajari, K. Kuruvinashetti, A. Komeili, and U. Sundararaj, "The emergence of ai-based wearable sensors for digital health technology: A review", *Sensors*, vol. 23, no. 23, p. 9498, 2023.

[20] P. Kostkova, *Grand challenges in digital health*, 2015.

[21] E. J. Emanuel and R. M. Wachter, "Artificial intelligence in health care: Will the value match the hype?", *Jama*, vol. 321, no. 23, pp. 2281–2282, 2019.

[22] S. Roy, T. Meena, and S.-J. Lim, "Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine", *Diagnostics*, vol. 12, no. 10, p. 2549, 2022.

[23] A. Trezza, A. Visibelli, B. Roncaglia, O. Spiga, and A. Santucci, "Unsupervised learning in precision medicine: Unlocking personalized healthcare through ai", *Applied Sciences*, vol. 13, no. 20, p. 9305, 2023. DOI: 10.3390/app13209305.

[24] A. Yakimovich, A. Beaugnon, Y. Huang, and E. Ozkirimli, "Labels in a haystack: Approaches beyond supervised learning in biomedical applications", *Patterns*, vol. 2, no. 12, 2021.

[25] S. D. Holcomb, W. K. Porter, S. V. Ault, G. Mao, and J. Wang, "Overview on deepmind and its alphago zero ai", in *Proceedings of the 2018 international conference on big data and education*, 2018, pp. 67–71.

[26] M. V. Koroteev, "Bert: A review of applications in natural language processing and understanding", *arXiv preprint arXiv:2103.11943*, 2021.

[27] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets", in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[30] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[31] I. Perez-Pozuelo, B. Zhai, J. Palotti, *et al.*, "The future of sleep health: A data-driven revolution in sleep science and medicine", *NPJ digital medicine*, vol. 3, no. 1, p. 42, 2020.

[32] 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.

[33] S. Biswal, J. Kulas, H. Sun, *et al.*, "Sleepnet: Automated sleep staging system via deep learning", *arXiv preprint arXiv:1707.08262*, 2017.

[34] R. Haidar, I. Koprinska, and B. Jeffries, "Sleep apnea event detection from nasal airflow using convolutional neural networks", in *International Conference on Neural Information Processing*, Springer, 2017, pp. 819–827.

[35] C. Loza, L. Colgin, d. B. M, *et al.*, "Deep Neural Dynamic Bayesian Networks Applied to EEG Sleep Spindles Modeling", English, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12905, pp. 550–560, 2021.

[36] S. Nikkonen, P. Somaskandhan, H. Korkalainen, *et al.*, "Multicentre sleep-stage scoring agreement in the sleep revolution project", *Journal of Sleep Research*, vol. 33, no. 1, e13956, 2024.

[37] T. M. Mitchell, "Machine learning", *McGraw-hill New York*, vol. 1, no. 9, 1997.

[38] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT Press Cambridge, 2016, vol. 1.

[39] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.", *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[40] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: An overview", in *Journal of physics: conference series*, IOP Publishing, vol. 1142, 2018, p. 012 012.

[41] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[42] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms", *Electronics*, vol. 12, no. 8, p. 1789, 2023.

[43] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare", *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1346–1352, 2022.

[44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[45] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[46] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An unsupervised machine learning algorithms: Comprehensive review", *International Journal of Computing and Digital Systems*, 2023.

[47] M. Abukmeil, S. Ferrari, A. Genovese, V. Piuri, and F. Scotti, "A survey of unsupervised generative models for exploratory data analysis and representation learning", *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–40, 2021.

[48] M. J. Sateia, "International classification of sleep disorders", *Chest*, vol. 146, no. 5, pp. 1387–1394, 2014.

[49] H. Rajaguru, R. Karthikamani, H. R, B. C.G, P. C, and D. D, "Performance Analysis of the Classifier in the Classification of Normal-Sleep and Seizure from EEG Signal", English, *Proceedings - 2nd International Conference on Smart Technologies, Communication and Robotics 2022, STCR 2022*, 2022.

[50] L. Anghel, A. Ciubară, A. Nechita, *et al.*, "Sleep disorders associated with neurodegenerative diseases", *Diagnostics*, vol. 13, no. 18, p. 2898, 2023.

[51] K. Ramar and E. J. Olson, "Management of common sleep disorders", *American family physician*, vol. 88, no. 4, pp. 231–238, 2013.

[52] M. Óskarsdóttir, A. Islind, E. August, E. Arnardóttir, F. Patou, and A. Maier, "Importance of Getting Enough Sleep and Daily Activity Data to Assess Variability: Longitudinal Observational Study", English, *JMIR Formative Research*, vol. 6, no. 2, 2022.

[53] M. A. Carskadon, W. C. Dement, *et al.*, "Normal human sleep: An overview", *Principles and practice of sleep medicine*, vol. 4, no. 1, pp. 13–23, 2005.

[54] S. Chokroverty *et al.*, "Overview of sleep & sleep disorders", *Indian J Med Res*, vol. 131, no. 2, pp. 126–140, 2010.

[55] D.-J. Dijk, "Regulation and functional correlates of slow wave sleep", *Journal of Clinical Sleep Medicine*, vol. 5, no. 2 suppl, S6–S15, 2009.

[56] P. A. Bryant, J. Trinder, and N. Curtis, "Sick and tired: Does sleep have a vital role in the immune system?", *Nature Reviews Immunology*, vol. 4, no. 6, pp. 457–467, 2004.

[57] W. Huang, K. M. Ramsey, B. Marcheva, J. Bass, *et al.*, "Circadian rhythms, sleep, and metabolism", *The Journal of clinical investigation*, vol. 121, no. 6, pp. 2133–2141, 2011.

[58] R. L. Sack, D. Auckley, R. R. Auger, *et al.*, "Circadian rhythm sleep disorders: Part i, basic principles, shift work and jet lag disorders", *Sleep*, vol. 30, no. 11, pp. 1460–1483, 2007.

[59] W. H. Walker, J. C. Walton, A. C. DeVries, and R. J. Nelson, "Circadian rhythm disruption and mental health", *Translational psychiatry*, vol. 10, no. 1, p. 28, 2020.

[60] E. J. Stepanski and J. K. Wyatt, "Use of sleep hygiene in the treatment of insomnia", *Sleep medicine reviews*, vol. 7, no. 3, pp. 215–225, 2003.

[61] D. J. Buysse, "Sleep health: Can we define it? does it matter?", *Sleep*, vol. 37, no. 1, pp. 9–17, 2014.

[62] J. Orzeł-Gryglewska, "Consequences of sleep deprivation.", *International journal of occupational medicine and environmental health*, 2010.

[63] W. D. Killgore, "Effects of sleep deprivation on cognition", *Progress in brain research*, vol. 185, pp. 105–129, 2010.

[64] P. Maruff, M. G. Falleti, A. Collie, D. Darby, and M. McStephen, "Fatigue-related impairment in the speed, accuracy and variability of psychomotor performance: Comparison with blood alcohol levels", *Journal of sleep research*, vol. 14, no. 1, pp. 21–27, 2005.

[65] A. D. Krystal and J. D. Edinger, "Measuring sleep quality", *Sleep medicine*, vol. 9, S10–S17, 2008.

[66] L. Zhang and Z.-X. Zhao, "Objective and subjective measures for sleep disorders", *Neuroscience bulletin*, vol. 23, no. 4, p. 236, 2007.

[67] M. Bruyneel, C. Sanida, G. Art, *et al.*, "Sleep efficiency during sleep studies: Results of a prospective study comparing home-based and in-hospital polysomnography", *Journal of sleep research*, vol. 20, no. 1pt2, pp. 201–206, 2011.

[68] A. Sadeh, "The role and validity of actigraphy in sleep medicine: An update", *Sleep Medicine Reviews*, vol. 15, no. 4, pp. 259–267, 2011.

[69] F. C. Baker, S. Maloney, and H. S. Driver, "A comparison of subjective estimates of sleep with objective polysomnographic data in healthy men and women", *Journal of psychosomatic research*, vol. 47, no. 4, pp. 335–341, 1999.

[70] Y. S. Bin, "Is sleep quality more important than sleep duration for public health?", *Sleep*, vol. 39, no. 9, pp. 1629–1630, 2016.

[71] M. P. Hoevenaar-Blom, A. M. Spijkerman, D. Kromhout, J. F. van den Berg, and W. Verschuren, "Sleep duration and sleep quality in relation to 12-year cardiovascular disease incidence: The morgen study", *Sleep*, vol. 34, no. 11, pp. 1487–1492, 2011.

[72] L. R. Pinto, M. C. R. Pinto, L. I. Goulart, *et al.*, "Sleep perception in insomniacs, sleep-disordered breathing patients, and healthy volunteers–an important biologic parameter of sleep", *Sleep Medicine*, vol. 10, no. 8, pp. 865–868, 2009.

[73] T. Åkerstedt, J. Schwarz, G. Gruber, E. Lindberg, and J. Theorell-Haglöw, "The relation between polysomnography and subjective sleep and its dependence on age–poor sleep may become good sleep", *Journal of Sleep Research*, vol. 25, no. 5, pp. 565–570, 2016.

[74] M. K. Means, J. D. Edinger, D. M. Glenn, and A. I. Fins, "Accuracy of sleep perceptions among insomnia sufferers and normal sleepers", *Sleep medicine*, vol. 4, no. 4, pp. 285–296, 2003.

[75] E. Björnsdóttir, C. Janson, T. Gíslason, *et al.*, "Insomnia in untreated sleep apnea patients compared to controls", *Journal of sleep research*, vol. 21, no. 2, pp. 131–138, 2012.

[76] D. Riemann, C. Baglioni, C. Bassetti, *et al.*, "European guideline for the diagnosis and treatment of insomnia", *Journal of sleep research*, vol. 26, no. 6, pp. 675–700, 2017.

[77] B. F. Boeve, M. Silber, C. Saper, *et al.*, "Pathophysiology of rem sleep behaviour disorder and relevance to neurodegenerative disease", *Brain*, vol. 130, no. 11, pp. 2770–2788, 2007.

[78] M. Manconi, D. Garcia-Borreguero, B. Schormair, *et al.*, "Restless legs syndrome", *Nature reviews Disease primers*, vol. 7, no. 1, p. 80, 2021.

[79] M. Hornyak, B. Feige, D. Riemann, and U. Voderholzer, "Periodic leg movements in sleep and periodic limb movement disorder: Prevalence, clinical significance and treatment", *Sleep medicine reviews*, vol. 10, no. 3, pp. 169–177, 2006.

[80] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults", *American journal of epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.

[81] N. M. Punjabi, "The epidemiology of adult obstructive sleep apnea", *Proceedings of the American Thoracic Society*, vol. 5, no. 2, pp. 136–143, 2008.

[82] P. J. Strollo Jr and R. M. Rogers, "Obstructive sleep apnea", *New England Journal of Medicine*, vol. 334, no. 2, pp. 99–104, 1996.

[83] 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", *The Lancet Respiratory Medicine*, vol. 7, no. 8, pp. 687–698, 2019.

[84] D. A. Pevernagie, B. Gnidovec-Strazisar, L. Grote, *et al.*, "On the rise and fall of the apnea- hypopnea index: A historical review and critical appraisal", *Journal of sleep research*, vol. 29, no. 4, e13066, 2020.

[85] L. Prochnow, S. Zimmermann, and T. Penzel, "Predictors of obstructive sleep apnea", *Somnologie*, vol. 20, no. 2, pp. 113–118, 2016.

[86] P. Lévy, J.-L. Pépin, P. Mayer, B. Wuyam, and D. Veale, "Management of simple snoring, upper airway resistance syndrome, and moderate sleep apnea syndrome", *Sleep*, vol. 19, no. suppl_9, S101–S110, 1996.

[87] N. Mcardle, G. Devereux, H. Heidarnejad, H. M. Engleman, T. W. Mackay, and N. J. Douglas, "Long-term use of cpap therapy for sleep apnea/hypopnea syndrome", *American journal of respiratory and critical care medicine*, vol. 159, no. 4, pp. 1108–1114, 1999.

[88] K. Y. Fridgeirsdottir, C. J. Murphy, A. S. Islind, *et al.*, "Effects of exercise and a lifestyle app on sleep-disordered breathing, physical health, and quality of life", *ERJ Open Research*, 2024.

[89] C. López-Padrós, N. Salord, C. Alves, *et al.*, "Effectiveness of an intensive weight-loss program for severe osa in patients undergoing cpap treatment: A randomized controlled trial", *Journal of Clinical Sleep Medicine*, vol. 16, no. 4, pp. 503–514, 2020.

[90] M. Camacho, V. Certal, J. Abdullatif, *et al.*, "Myofunctional therapy to treat obstructive sleep apnea: A systematic review and meta-analysis", *Sleep*, vol. 38, no. 5, pp. 669–675, 2015.

[91] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling", *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014.

[92] K. Šušmáková, "Human sleep and sleep eeg", *Measurement science review*, vol. 4, no. 2, pp. 59–74, 2004.

[93] P. Halász, "K-complex, a reactive eeg graphoelement of nrem sleep: An old chap in a new garment", *Sleep medicine reviews*, vol. 9, no. 5, pp. 391–412, 2005.

[94] L. C. Markun and A. Sampat, "Clinician-focused overview and developments in polysomnography", *Current sleep medicine reports*, pp. 1–13, 2020.

[95] K. Montazeri, S. A. Jonsson, J. S. Agustsson, M. Serwatko, T. Gislason, and E. S. Arnardottir, "The design of rip belts impacts the reliability and quality of the measured respiratory signals", *Sleep and Breathing*, pp. 1–7, 2021.

[96] M. Hirshkowitz, "Principles and practice of sleep medicine (sixth edition)", in M. Kryger, T. Roth, and W. C. Dement, Eds., Sixth Edition, Elsevier, 2017, 1564–1566.e3, ISBN: 978-0-323-24288-2.

[97] S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García, "SpO2 based sleep apnea detection using deep learning", English, *INES 2017 - IEEE 21st International Conference on Intelligent Engineering Systems, Proceedings*, vol. 2017, pp. 91–96, 2017.

[98] J. V. Marcos, R. Hornero, D. Alvarez, F. del Campo, M. López, and C. Zamarrón, "Radial basis function classifiers to help in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry.", eng, *Medical & biological engineering & computing*, vol. 46, no. 4, pp. 323–332, Apr. 2008.

[99] Z. Li, M. Arvaneh, H. Elphick, R. Kingshott, and L. Mihaylova, "A dirichlet process mixture model for autonomous sleep apnea detection using oxygen saturation data", English, *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020*, 2020.

[100] M. A. Almarshad, S. Al-Ahmadi, M. S. Islam, A. S. Ba-Hammam, and A. Soudani, "Adoption of Transformer Neural Network to Improve the Diagnostic Performance of Oximetry for Obstructive Sleep Apnea.", eng, *Sensors (Basel, Switzerland)*, vol. 23, no. 18, Sep. 2023.

[101] M. Takahashi, N. Sugahara, and M. Shibata, "Towards detecting morning surge from sleep self-evaluations", English, *Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2020*, pp. 1–6, 2020.

[102] N. Cooray, Z. Li, J. Wang, *et al.*, "Automated Movement Detection with Dirichlet Process Mixture Models and Electromyography", English, *2022 25th International Conference on Information Fusion, FUSION 2022*, 2022.

[103] M. Shokrollahi and S. Krishnan, "Sleep EMG analysis using sparse signal representation and classification.", eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2012, pp. 3480–3483, 2012.

[104] M. Daley, K. Diaz, H. Posada-Quintero, Y. Kong, K. Chon, and J. Bolkhovsky, "Archetypal physiological responses to prolonged wakefulness", English, *Biomedical Signal Processing and Control*, vol. 74, 2022.

[105] M. Óskarsdóttir, A. S. Islind, E. August, E. S. Arnardóttir, F. Patou, A. M. Maier, *et al.*, "Importance of getting enough sleep and daily activity data to assess variability: Longitudinal observational study", *JMIR Formative Research*, vol. 6, no. 2, e31807, 2022.

[106] T. Q. Le, C. Cheng, A. Sangasoongsong, W. Wongdhamma, and S. T. S. Bukkapatnam, "Wireless Wearable Multisensory Suite and Real-Time Prediction of Obstructive Sleep Apnea Episodes.", eng, *IEEE journal of translational engineering in health and medicine*, vol. 1, p. 2 700 109, 2013.

[107] T. Bajkowski, N. Marchal, J. Saied-Walker, *et al.*, "Cohort Discovery from Bed Sensor Data with Fuzzy Evidence Accumulation Clustering", English, *IEEE International Conference on Fuzzy Systems*, 2023.

[108] C. Köhler, A. Bartschke, D. Fuerstenau, T. Schaaf, and E. Salgado-Baez, "The value of smartwatches in the healthcare sector: Monitoring, nudging, and predicting (preprint)", *Journal of Medical Internet Research*, Mar. 2024.

[109] V. Ramnath and S. Katkoori, "A Smart IoT System for Continuous Sleep State Monitoring", English, *Midwest Symposium on Circuits and Systems*, vol. 2020, pp. 241–244, 2020.

[110] Q. Pan, D. Brulin, E. Campo, and P. S, "Home sleep monitoring based on wrist movement data processing", English, *Procedia Computer Science*, vol. 183, pp. 696–705, 2021.

[111] A. Sebastian, P. Cistulli, G. Cohen, and P. d. Chazal, "A Preliminary Study of the Automatic Classification of the Site of Airway Collapse in OSA patients Using Snoring Signals.", eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2019, pp. 1592–1595, Jul. 2019.

[112] I. Takao, K. Nishio, T. Kaburagi, S. Kumagai, T. Matsumoto, and Y. Kurihara, "A Home Sleep Apnea State Monitoring System using a Stacked Autoencoder", English, *Proceedings of IEEE Sensors*, vol. 2019, 2019.

[113] A. Alfeo, P. Barsocchi, M. Cimino, D. La Rosa, F. Palumbo, and G. Vaglini, "Sleep behavior assessment via smartwatch and stigmergic receptive fields", English, *Personal and Ubiquitous Computing*, vol. 22, no. 2, pp. 227–243, 2018.

[114] S. A. A. Massar, X. Y. Chua, C. S. Soon, *et al.*, "Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data.", eng, *NPJ digital medicine*, vol. 4, no. 1, p. 90, Jun. 2021.

[115] S. Bhatlawande and S. Kulkarni, "Residential Monitoring System for Classification and Recognition of Sleeping Posture", English, *2022 2nd International Conference on Intelligent Technologies, CONIT 2022*, 2022.

[116] A. Heinrich, X. Zhao, and G. De Haan, "Multi-distance motion vector clustering algorithm for video-based sleep analysis", English, *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services, Healthcom 2013*, pp. 223–227, 2013.

[117] N. Koolen, O. Decroupet, A. Dereymaeker, *et al.*, "Automated respiration detection from neonatal video data", English, *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings*, vol. 2, pp. 164–169, 2015.

[118] K. Zhu, M. Li, S. Akbarian, M. Hafezi, A. Yadollahi, and B. Taati, "Vision-Based Heart and Respiratory Rate Monitoring During Sleep - A Validation Study for the Population at Risk of Sleep Apnea.", eng, *IEEE journal of translational engineering in health and medicine*, vol. 7, p. 1 900 708, 2019.

[119] F. Barata, P. Tinschert, F. Rassouli, *et al.*, "Automatic Recognition, Segmentation, and Sex Assignment of Nocturnal Asthmatic Coughs and Cough Epochs in Smartphone Audio Recordings: Observational Field Study.", eng, *Journal of medical Internet research*, vol. 22, no. 7, e18082, Jul. 2020.

[120] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snore-GANs: Improving Automatic Snore Sound Classification with Synthesized Data", English, *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 300–310, 2020.

[121] R. J. Beeton, I. Wells, P. Ebden, H. B. Whittet, and J. Clarke, "Snore site discrimination using statistical moments of free field snoring sounds recorded during sleep nasendoscopy.", eng, *Physiological measurement*, vol. 28, no. 10, pp. 1225–1236, Oct. 2007.

[122] J. Blanco, L. Hernández, R. Fernández, and D. Ramos, "Improving Automatic Detection of Obstructive Sleep Apnea Through Nonlinear Analysis of Sustained Speech", English, *Cognitive Computation*, vol. 5, no. 4, pp. 458–472, 2013.

[123] J.-A. Gómez-García, J.-L. Blanco-Murillo, J.-I. Godino-Llorente, L. Hernández Gómez, and G. Castellanos-Domínguez, "GMM-based classifiers for the automatic detection of Obstructive Sleep Apnea", English, *BIOSIGNALS 2013 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, pp. 364–367, 2013.

[124] R. Fernández, J. Blanco, L. Hernández, E. López, J. Alcázar, and D. Toledano, "Severe APNOEA detection using speaker recognition techniques", English, *BIOSIGNALS 2009 - Proceedings of the 2nd International Conference on Bio-Inspired Systems and Signal Processing*, pp. 124–130, 2009.

[125] N. P. Walsh, D. S. Kashi, J. P. Edwards, *et al.*, "Good perceived sleep quality protects against the raised risk of respiratory infection during sleep restriction in young adults", *Sleep*, 2022.

[126] C. E. Carney, D. J. Buysse, S. Ancoli-Israel, *et al.*, "The consensus sleep diary: Standardizing prospective sleep self-monitoring", *Sleep*, vol. 35, no. 2, pp. 287–302, 2012.

[127] L. Schmitz, B. F. Sveinbjarnarson, G. N. Gunnarsson, *et al.*, "Towards a digital sleep diary standard", *Proceedings of the Americas Conference on Information Systems (AMCIS), Minneappolis, August 9-13*, 2022.

[128] R. B. Berry, R. Brooks, C. E. Gamaldo, *et al.*, "AASM Manual for the Scoring of Sleep and Associated Events", American Academy of Sleep Medicine, Amer. Acad. Sleep Med., Darien, IL, USA, Tech. Rep., 2018, Version 2.5.

[129] E. S. Arnardottir, A. S. Islind, M. Óskarsdottir, *et al.*, "The sleep revolution project: The concept and objectives", *Journal of sleep research*, vol. 31, no. 4, e13630, 2022.

[130] B. Kemp, A. Värri, A. C. Rosa, K. D. Nielsen, and J. Gade, "A simple format for exchange of digitized polygraphic recordings", *Electroencephalography and clinical neurophysiology*, vol. 82, no. 5, pp. 391–393, 1992.

[131] U. J. Magalang, N.-H. Chen, P. A. Cistulli, *et al.*, "Agreement in the scoring of respiratory events and sleep among international sleep centers", *Sleep*, vol. 36, no. 4, pp. 591–596, 2013.

[132] T. Penzel, X. Zhang, and I. Fietze, "Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules", *Journal of Clinical Sleep Medicine*, vol. 9, no. 1, pp. 89–91, 2013.

[133] S. Sarker, S. Chatterjee, X. Xiao, and A. Elbanna, "The sociotechnical axis of cohesion for the is discipline", *MIS quarterly*, vol. 43, no. 3, 695–A5, 2019.

[134] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review", *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[135] A. Saxena, M. Prasad, A. Gupta, *et al.*, "A review of clustering techniques and developments", *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[136] L. Rokach and O. Maimon, "Clustering methods", *Data mining and knowledge discovery handbook*, pp. 321–352, 2005.

[137] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm", *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[138] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California press, vol. 5, 1967, pp. 281–298.

[139] T. Wongsirichot, A. Hanskunatai, B. V, H. D.-S, and P. P, "A comparative investigation of PSG signal patterns to classify sleep disorders using machine learning techniques", English, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9225, pp. 510–521, 2015.

[140] S. Mai, S. Amer-Yahia, S. Bailly, *et al.*, "Evolutionary Active Constrained Clustering for Obstructive Sleep Apnea Analysis", English, *Data Science and Engineering*, vol. 3, no. 4, pp. 359–378, 2018.

[141] E.-Y. Ma, J.-W. Kim, Y. Lee, S.-W. Cho, H. Kim, and J. Kim, "Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea", English, *Scientific Reports*, vol. 11, no. 1, 2021.

[142] P. Loliencar and G. Heo, "Phenotyping OSA: A time series analysis using fuzzy clustering and persistent homology", English, *International Journal of Approximate Reasoning*, vol. 142, pp. 178–195, 2022.

[143] W.-J. Cheng, E. Finnsson, E. Arnardóttir, J. S. Ágústsson, S. A. Sands, and L.-W. Hang, "Relationship between symptom profiles and endotypes among patients with obstructive sleep apnea: A latent class analysis", *Annals of the American Thoracic Society*, vol. 20, no. 9, pp. 1337–1344, 2023.

[144] S. Park, S. W. Lee, S. Han, and M. Cha, "Clustering Insomnia Patterns by Data From Wearable Devices: Algorithm Development and Validation Study.", eng, *JMIR mHealth and uHealth*, vol. 7, no. 12, e14473, Dec. 2019.

[145] X. Wang and Z. Xu, "Automatic Sleep Staging based on Curriculum Learning Approach", ICBIP '19, pp. 1–6, 2019.

[146] M. Bagci, T. Nguyen, and Y. Ozturk, "Ambient Sleep Quality Analysis with a Machine Learning Model", English, *ICASSPW 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing Workshops, Proceedings*, 2023.

[147] A. Yadollahi and Z. Moussavi, "Formant analysis of breath and snore sounds.", eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2009, pp. 2563–2566, 2009.

[148] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm", *Computers & geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[149] Y. El-Manzalawy, O. Buxton, V. Honavar, *et al.*, "Sleep/wake state prediction and sleep parameter estimation using unsupervised classification via clustering", English, *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, vol. 2017, pp. 718–723, 2017.

[150] R.-S. Hsiao, T.-X. Chen, M. A. Bitew, C.-H. Kao, and T.-Y. Li, "Sleeping posture recognition using fuzzy c-means algorithm.", eng, *Biomedical engineering online*, vol. 17, p. 157, Nov. 2018.

[151] P. Boyraz, M. Acar, and D. Kerr, "Multi-sensor driver drowsiness monitoring", English, *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 222, no. 11, pp. 1857–1878, 2008.

[152] S. C. Johnson, "Hierarchical clustering schemes", *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[153] H. Escola, E. Poiseau, M. Jobert, and P. Gaillard, "Classification using distance-based segmentation-application to the analysis of EEG signals", English, *Pattern Recognition Letters*, vol. 12, no. 6, pp. 327–333, 1991.

[154] Y. Amor, L. Rejeb, R. Ferjeni, L. Ben Said, and M. Ben Cheikh, "Hierarchical Multi-agent System for Sleep Stages Classification", English, *International Journal on Artificial Intelligence Tools*, vol. 31, no. 5, 2022.

[155] V. Gerla, M. Murgas, A. Mladek, *et al.*, "Hybrid hierarchical clustering algorithm used for large datasets: A pilot study on long-term sleep data", English, *IFMBE Proceedings*, vol. 66, pp. 3–7, 2018.

[156] T. Lajnef, S. Chaibi, P. Ruby, *et al.*, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines.", eng, *Journal of neuroscience methods*, vol. 250, pp. 94–105, Jul. 2015.

[157] J. Paalasmaa, H. Toivonen, and M. Partinen, "Adaptive Heartbeat Modeling for Beat-to-Beat Heart Rate Measurement in Ballistocardiograms.", eng, *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1945–1952, Nov. 2015.

[158] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering", *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[159] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: Why and how you should (still) use dbscan", *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[160] Y. Wang, I. Azimi, M. Feli, A. M. Rahmani, and P. Liljeberg, "Personalized Graph Attention Network for Multivariate Time-series Change Analysis: A Case Study on Long-term Maternal Monitoring", SAC '23, pp. 593–598, 2023.

[161] E. Nasibov, M. Ozgören, G. Ulutagay, A. Oniz, and S. Kocaaslan, "On the analysis of BIS stage epochs via fuzzy clustering.", eng, *Biomedizinische Technik. Biomedical engineering*, vol. 55, no. 3, pp. 147–153, Jun. 2010.

[162] Y. Yu, B. Wang, J. Jin, X. Wang, L. Q, and W. L, "Automatic Sleep Stage Classification by a Density - Distance-Based K - means Clustering Algorithm with Amendments", English, *Proceedings - 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2019*, 2019.

[163] L. Sepulveda-Cano, E. Gil, P. Laguna, and G. Castellanos-Dominguez, "Selection of nonstationary dynamic features for obstructive sleep apnoea detection in children", English, *Eurasip Journal on Advances in Signal Processing*, vol. 2011, 2011.

[164] R. Wei, X. Zhang, J. Wang, and X. Dang, "The research of sleep staging based on single-lead electrocardiogram and deep neural network.", eng, *Biomedical engineering letters*, vol. 8, no. 1, pp. 87–93, Feb. 2018.

[165] S. Kim, D. Lee, H. Kwak, and S. Lee, "Towards Domain-free Transformer for Generalized EEG Pre-training", English, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 1–1, 2024.

[166] S. S. D. P. Ayyagari, R. D. Jones, and S. J. Weddell, "Detection of microsleep states from the EEG: A comparison of feature reduction methods.", eng, *Medical & biological engineering & computing*, vol. 59, no. 7, pp. 1643–1657, Aug. 2021.

[167] R. Haidar, I. Koprinska, B. Jeffries, G. T, W. K.W, and L. M, "Feature learning and data compression of biosignals using convolutional autoencoders for sleep apnea detection", English, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11953, pp. 162–174, 2019.

[168] D. Dhongade, T. Rao, P. S, P. K, and J. S, "Classification of sleep disorders based on EEG signals by using feature extraction techniques with KNN classifier", English, *IEEE International Conference on Innovations in Green Energy and Healthcare Technologies - 2017, IGEHT 2017*, 2017.

[169] I. Jolliffe, *Principal Component Analysis*. Wiley Online Library, 2002.

[170] A. Azarbarzin and Z. M. K. Moussavi, "Automatic and unsupervised snore sound extraction from respiratory sound signals.", eng, *IEEE transactions on bio-medical engineering*, vol. 58, no. 5, pp. 1156–1162, May 2011.

[171] M. Zubair, U. M, R. Tripathy, M. Alhartomi, S. Alzahrani, and S. Ahamed, "Detection of Sleep Apnea from ECG Signals using Sliding Singular Spectrum based Sub-Pattern Principal Component Analysis", English, *IEEE Transactions on Artificial Intelligence*, pp. 1–10, 2023.

[172] L. M. Sepúlveda-Cano, E. Gil, P. Laguna, and G. Castellanos-Dominguez, "Sleep apnoea detection in children using PPG envelope-based dynamic features.", eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2011, pp. 1483–1486, 2011.

[173] A. Fisher, B. Caffo, B. Schwartz, and V. Zipunnikov, "Fast, Exact Bootstrap Principal Component Analysis for p > 1 million.", eng, *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 846–860, 2016.

[174] H. Mao, J. Widjaja, Y. Guo, J. Yin, and R. Vinjamuri, "Finding Robust Low Dimensional Features for Sleep Detection Using EEG Data", English, *2022 IEEE 2nd International Conference on Data Science and Computer Application, ICDSCA 2022*, pp. 42–45, 2022.

[175] C. Metzner, A. Schilling, M. Traxdorf, *et al.*, "Classification at the accuracy limit: Facing the problem of data ambiguity.", eng, *Scientific reports*, vol. 12, no. 1, p. 22 121, Dec. 2022.

[176] A. Hyvärinen, "Independent component analysis: Recent advances", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20 110 534, 2013.

[177] Q. Nguyen, T. Le, Q. Vu, *et al.*, "An Algorithm for Removing Artifacts in Polysomnography Signals", English, *IFMBE Proceedings*, vol. 85, pp. 1017–1031, 2022.

[178] R. Sekkal, F. Bereksi-Reguig, N. Dib, *et al.*, "An Approach to Detecting and Eliminating Artifacts from the Sleep EEG Signals", English, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12108, pp. 155–160, 2020.

[179] J.-H. Lee, S. Oh, F. A. Jolesz, H. Park, and S.-S. Yoo, "Application of independent component analysis for the data mining of simultaneous Eeg-fMRI: Preliminary experience on sleep onset.", eng, *The International journal of neuroscience*, vol. 119, no. 8, pp. 1118–1136, 2009.

[180] M. Crespo-Garcia, M. Atienza, and J. L. Cantero, "Muscle artifact removal from human sleep EEG by using independent component analysis.", eng, *Annals of biomedical engineering*, vol. 36, no. 3, pp. 467–475, Mar. 2008.

[181] M. Uchida, W. Chen, T. Nemoto, K. Kitamura, Y. Kanemitsu, and D. Wei, "An ICA approach to reject noise from pressure changes of pillow", English, *Proceedings - The Fourth International Conference on Computer and Information Technology (CIT 2004)*, pp. 916–921, 2004.

[182] Y. S. Lee, P. N. Pathirana, R. J. Evans, and C. L. Steinfort, "Separation of Doppler radar-based respiratory signatures.", eng, *Medical & biological engineering & computing*, vol. 54, no. 8, pp. 1169–1179, Aug. 2016.

[183] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview", *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 71–82, 2006.

[184] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, *et al.*, "Fast discovery of association rules.", *Advances in knowledge discovery and data mining*, vol. 12, no. 1, pp. 307–328, 1996.

[185] C. Borgelt, "Frequent item set mining", *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 2, no. 6, pp. 437–456, 2012.

[186] Z. Liang, B. Ploderer, M. Martell, T. Nishimura, M.-G. A, and U.-C. V, "A cloud-based intelligent computing system for contextual exploration on personal sleep-tracking data using association rule mining", English, *Communications in Computer and Information Science*, vol. 597, pp. 83–96, 2016.

[187] P. Laxminarayan, S. Alvarez, C. Ruiz, and M. Moonis, "Mining associations over human sleep time series", English, *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, pp. 323–325, 2005.

[188] J.-C. Kim and K. Chung, "Mining Based Time-Series Sleeping Pattern Analysis for Life Big-Data", English, *Wireless Personal Communications*, vol. 105, no. 2, pp. 475–489, 2019.

[189] Z. Liang, M. Martell, and T. Nishimura, "Mining hidden correlations between sleep and lifestyle factors from quantified-self data", English, *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 547–552, 2016.

[190] P. Laxminarayan, S. A. Alvarez, C. Ruiz, and M. Moonis, "Mining statistically significant associations for exploratory analysis of human sleep data.", eng, *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 10, no. 3, pp. 440–450, Jul. 2006.

[191] R. Abeysinghe and L. Cui, "Query-constraint-based mining of association rules for exploratory analysis of clinical datasets in the National Sleep Research Resource.", eng, *BMC medical informatics and decision making*, vol. 18, p. 58, Jul. 2018.

[192] M. R. Álvarez, P. Félix, and P. Cariñena, "Discovering metric temporal constraint networks on temporal databases.", eng, *Artificial intelligence in medicine*, vol. 58, no. 3, pp. 139–154, Jul. 2013.

[193] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks", *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[194] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview", *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[195] A. Salazar, L. Vergara, and G. Safont, "Generative adversarial networks and markov random fields for oversampling very small training sets", *Expert Systems with Applications*, vol. 163, p. 113 819, 2021.

[196] K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, S. Van Huffel, and M. De Vos, "Automated eeg sleep staging in the term-age baby using a generative modelling approach", *Journal of neural engineering*, vol. 15, no. 3, p. 036 004, 2018.

[197] C. A. Loza and J. C. Principe, "The generalized sleep spindles detector: A generative model approach on single-channel eegs", in *Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I 15*, Springer, 2019, pp. 127–138.

[198] C.-E. Kuo, T.-H. Lu, G.-T. Chen, and P.-Y. Liao, "Towards precision sleep medicine: Self-attention gan as an innovative data augmentation technique for developing personalized automatic sleep scoring classification", *Computers in Biology and Medicine*, vol. 148, p. 105 828, 2022.

[199] C. Loza, J. Principe, J. G, R. I, and C. A, "The Generalized Sleep Spindles Detector: A Generative Model Approach on Single-Channel EEGs", English, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11506, pp. 127–138, 2019.

[200] T. Takeda, O. Mizuno, and T. Tanaka, "Time-dependent sleep stage transition model based on heart rate variability.", eng, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2015, pp. 2343–2346, 2015.

[201] C. M. Fernandes, A. Mora, J. J. Merelo, F. Fernández, and A. Rosa, "Generating colored 2-dimensional representations of sleep EEG with the KANTS clustering algorithm", GECCO '12, pp. 435–442, 2012.

[202] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey", *IEEE transactions on knowledge and data engineering*, vol. 24, no. 5, pp. 823–839, 2010.

[203] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review", *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[204] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques", *Procedia Computer Science*, vol. 60, pp. 708–713, 2015.

[205] H. Wang, A. Li, T. Chen, J. Liu, and S. Y, "Study on behavioral risk for aging based on smart mattress monitoring data", English, *Procedia Computer Science*, vol. 221, pp. 1276–1283, 2023.

[206] A. Gasmi, V. Augusto, J. Faucheu, C. Morin, and X. Serpaggi, "Anomaly Detection in Sleep Habits Using Deep Learning", English, *IEEE International Conference on Automation Science and Engineering*, vol. 2023, 2023.

[207] J.-J. Chung and H.-J. Kim, "An automobile environment detection system based on deep neural network and its implementation using iot-enabled in-vehicle air quality sensors", English, *Sustainability (Switzerland)*, vol. 12, no. 6, 2020.

[208] K. Fujiwara, E. Abe, K. Kamata, *et al.*, "Heart Rate Variability-Based Driver Drowsiness Detection and Its Validation With EEG.", eng, *IEEE transactions on bio-medical engineering*, vol. 66, no. 6, pp. 1769–1778, Jun. 2019.

[209] A. Caroppo, A. Leone, G. Rescio, *et al.*, "Multi-sensor platform for detection of anomalies in human sleep patterns", English, *Lecture Notes in Electrical Engineering*, vol. 431, pp. 276–285, 2018.

[210] L. Rabiner and B. Juang, "An introduction to hidden markov models", *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[211] S. R. Eddy, "What is a hidden markov model?", *Nature biotechnology*, vol. 22, no. 10, pp. 1315–1316, 2004.

[212] J. Kohlmorgen, K.-R. Müllerc, and K. Pawelzik, "Analysis of drifting dynamics with neural network hidden markov models", English, *Advances in Neural Information Processing Systems*, pp. 735–741, 1998.

[213] S. Alvarez and C. Ruiz, "Collective probabilistic dynamical modeling of sleep stage transitions", English, *BIOSIGNALS 2013 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, pp. 209–214, 2013.

[214] F. Mendonça, S. Mostafa, F. Morgado-Dias, and A. Ravelo-García, "Cyclic alternating pattern estimation based on a probabilistic model over an EEG signal", English, *Biomedical Signal Processing and Control*, vol. 62, 2020.

[215] E. Eldele, M. Ragab, Z. Chen, *et al.*, "ADAST: Attentive Cross-Domain EEG-Based Sleep Staging Framework With Iterative Self-Training", English, *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 1, pp. 210–221, 2023.

[216] D.-R. Gao, J. Li, M.-Q. Wang, L.-T. Wang, and Y.-Q. Zhang, "Automatic sleep staging of single-channel EEG based on domain adversarial neural networks and domain self-attention.", eng, *Frontiers in neuroscience*, vol. 17, p. 1 143 495, 2023.

[217] Z. He, M. Tang, P. Wang, *et al.*, "Cross-scenario automatic sleep stage classification using transfer learning and single-channel EEG", English, *Biomedical Signal Processing and Control*, vol. 81, 2023.

[218] E. R. M. Heremans, H. Phan, P. Borzée, B. Buyse, D. Testelmans, and M. De Vos, "From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging.", eng, *Journal of neural engineering*, vol. 19, no. 3, Jun. 2022.

[219] W. Qu, C.-H. Kao, H. Hong, *et al.*, "Single-channel EEG based insomnia detection with domain adaptation", English, *Computers in Biology and Medicine*, vol. 139, 2021.

[220] C. Yoo, H. Lee, and J.-W. Kang, "Transferring Structured Knowledge in Unsupervised Domain Adaptation of a Sleep Staging Network", English, *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1273–1284, 2022.

[221] Y. Luo, Y. Zheng, H. Shao, L. Zhang, and L. Li, "TUDAMatch: Time-Series Unsupervised Domain Adaptation for Automatic Sleep Staging", English, *International IEEE/EMBS Conference on Neural Engineering, NER*, vol. 2023, 2023.

[222] R. Zhao, Y. Xia, and Y. Zhang, "Unsupervised sleep staging system based on domain adaptation", English, *Biomedical Signal Processing and Control*, vol. 69, 2021.

[223] J. Fan, H. Zhu, X. Jiang, *et al.*, "Unsupervised Domain Adaptation by Statistics Alignment for Deep Sleep Staging Networks", English, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 205–216, 2022.

[224] G. Bazoukis, S. C. Bollepalli, C. T. Chung, *et al.*, "Application of artificial intelligence in the diagnosis of sleep apnea", *Journal of Clinical Sleep Medicine*, vol. 19, no. 7, pp. 1337–1363, 2023.

[225] X. Li, Y. Gong, X. Jin, and P. Shang, "Sleep posture recognition based on machine learning: A systematic review", *Pervasive and Mobile Computing*, vol. 90, p. 101 752, 2023.

[226] H. Alsolai, S. Qureshi, S. M. Z. Iqbal, *et al.*, "A systematic review of literature on automated sleep scoring", *IEEE Access*, vol. 10, pp. 79 419–79 443, 2022.

[227] G. C. Gutiérrez-Tobal, D. Álvarez, L. Kheirandish-Gozal, F. Del Campo, D. Gozal, and R. Hornero, "Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis", *Pediatric pulmonology*, vol. 57, no. 8, pp. 1931–1943, 2022.

[228] R. K. Yin, *Case study research: Design and methods*. sage, 2009, vol. 5.

[229] M. T. Pham, A. Rajić, J. D. Greig, J. M. Sargeant, A. Papadopoulos, and S. A. McEwen, "A scoping review of scoping reviews: Advancing the approach and enhancing the consistency", *Research synthesis methods*, vol. 5, no. 4, pp. 371–385, 2014.

[230] A. C. Tricco, E. Lillie, W. Zarin, *et al.*, "Prisma extension for scoping reviews (prisma-scr): Checklist and explanation", *Annals of internal medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[231] L. Biedebach, *Unsupervised machine learning in sleep research - a scoping review protocol*, Jul. 2024. [Online]. Available: `osf.io/42zrb`.

[232] G. N. Pires, E. S. Arnardóttir, J. M. Saavedra, S. Tufik, and W. T. McNicholas, "Search filters for systematic reviews and meta-analyses in sleep medicine", *Sleep Medicine*, 2025.

[233] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space", *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[234] Z. Shahid, S. Saguna, and C. Ahlund, "Recognizing Long-term Sleep Behaviour Change using Clustering for Elderly in Smart Homes", English, *ISC2 2022 - 8th IEEE International Smart Cities Conference*, 2022.

[235] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical ai", *Nature medicine*, vol. 28, no. 9, pp. 1773–1784, 2022.

[236] S. Fenton, S. Low, K. Abrams, and K. Butler-Henderson, "Health information management: Changing with time", *Yearbook of medical informatics*, vol. 26, no. 01, pp. 72–77, 2017.

[237] R. Bouderhem, "Privacy and regulatory issues in wearable health technology", *Engineering Proceedings*, vol. 58, no. 1, p. 87, 2023.

[238] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges", *PLoS medicine*, vol. 15, no. 11, e1002689, 2018.

[239] S. Canali, V. Schiaffonati, and A. Aliverti, "Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness", *PLOS Digital Health*, vol. 1, no. 10, e0000104, 2022.

[240] M. E. Salwei and P. Carayon, "A sociotechnical systems framework for the application of artificial intelligence in health care delivery", *Journal of cognitive engineering and decision making*, vol. 16, no. 4, pp. 194–206, 2022.

[241] T. J. Loftus, B. Shickel, J. A. Balch, *et al.*, "Phenotype clustering in health care: A narrative review for clinicians", *Frontiers in artificial intelligence*, vol. 5, p. 842 306, 2022.

[242] A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, *et al.*, "Machine learning (ml) in medicine: Review, applications, and challenges", *Mathematics*, vol. 9, no. 22, p. 2970, 2021.

[243] A. Smiti, "When machine learning meets medical world: Current status and future challenges", *Computer Science Review*, vol. 37, p. 100 280, 2020.

[244] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS quarterly*, pp. 319–340, 1989.

[245] S. I. Lambert, M. Madi, S. Sopka, *et al.*, "An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals", *NPJ Digital Medicine*, vol. 6, no. 1, p. 111, 2023.

[246] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey", *Computers in Biology and Medicine*, vol. 149, p. 106 043, 2022.

[247] Y. Xie, G. Gao, and X. Chen, "Outlining the design space of explainable intelligent systems for medical diagnosis", *arXiv preprint arXiv:1902.06019*, 2019.

[248] A. Shaban-Nejad, M. Michalowski, J. S. Brownstein, and D. L. Buckeridge, "Guest editorial explainable ai: Towards fairness, accountability, transparency and trust in healthcare", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2374–2375, 2021.

[249] M. Kahng, N. Thorat, D. H. P. Chau, F. B. Viégas, and M. Wattenberg, "GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation", *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 310–320, 2019, ISSN: 19410506. arXiv: 1809.01587.

[250] E. Werner, J. N. Clark, A. Hepburn, *et al.*, "Explainable hierarchical clustering for patient subtyping and risk prediction", *Experimental Biology and Medicine*, vol. 248, no. 24, pp. 2547–2559, 2023.

[251] S. S. Band, A. Yarahmadi, C.-C. Hsu, *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods", *Informatics in Medicine Unlocked*, vol. 40, p. 101 286, 2023.

[252] D. Hazra and Y.-C. Byun, "Synsiggan: Generative adversarial networks for synthetic biomedical signal generation", *Biology*, vol. 9, no. 12, p. 441, 2020.

[253] A. Sharma, A. Lysenko, S. Jia, K. A. Boroevich, and T. Tsunoda, "Advances in ai and machine learning for predictive medicine", *Journal of Human Genetics*, vol. 69, no. 10, pp. 487–497, 2024.

[254] H. Guan, L. Wang, and M. Liu, "Multi-source domain adaptation via optimal transport for brain dementia identification", in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2021, pp. 1514–1517.

[255] U. Lundberg, "Stress, subjective and objective health", *International Journal of Social Welfare*, vol. 15, S41–S48, 2006.

[256] P. D. Cleary, *Subjective and objective measures of health: Which is better when?*, 1997.

[257] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records", *Scientific reports*, vol. 6, no. 1, p. 26 094, 2016.

Appendix A

Publication I

SLEEP, 2024, pp. 1–38

doi: xxx
Advance Access Publication Date: Day Month Year

Unsupervised Machine Learning in Sleep Research: A Scoping Review

Luka Biedebach,[1,2,*] Daniela Ferreira-Santos,[3,4] Marie-Ange Stefanos,[5] Alva Lindhagen,[6] Gabriel Natan Pires,[7] Erna Sif Arnardóttir[1,2] and Anna Sigridur Islind[1]

[1]Department of Computer Science, Reykjavik University, Iceland, [2]Reykjavik University Sleep Institute, Reykjavik University, Reykjavik, Iceland, [3]Department of Medical Physics, University of Eastern Finland, Kuopio, Finland, [4]INESC TEC, Universidade do Porto, Porto, Portugal, [5]Computer Science Department, Université Paris-Cité, Paris, France, [6]Department of Computer Science, Umeå University, Umeå, Sweden and [7]Departmento de Psicobiologia, Universidade Federal de São Paulo, São Paulo, Brazil
*Corresponding author: Luka Biedebach, Department of Computer Science, Reykjavik University, Menntavegur 1, 102 Reykjavik, Iceland. Email: lukab@ru.is

Abstract

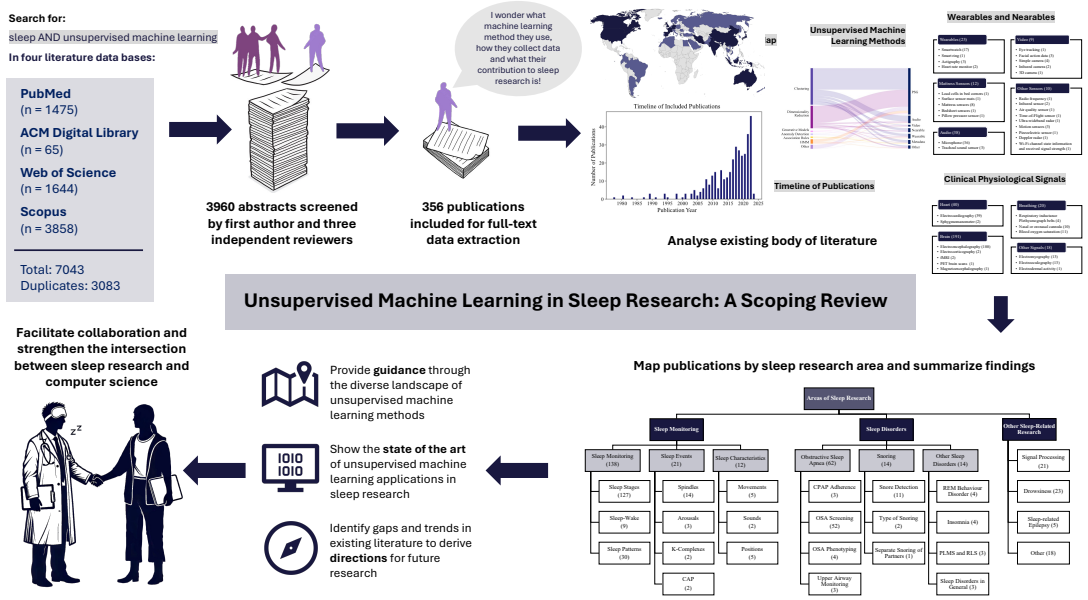
Study Objectives: Unsupervised machine learning —an approach that identifies patterns and structures within data without relying on manual labeling— has demonstrated remarkable success in various domains of sleep research. This underscores the broader utility of machine learning, suggesting that its capabilities extend beyond current applications and warrant further exploration for novel insights in sleep studies, focusing specifically on unsupervised machine learning.
Methods: This paper outlines a scoping review conducted according to the PRISMA guidelines for scoping reviews. A comprehensive search covering various search terms focusing on the intersection between unsupervised machine learning and sleep led to 3960 publications. After screening all titles and abstracts with two independent reviewers, ultimately, 356 publications were included in the full-text review. The data extracted from the full-texts included information about the machine learning methods and types of sleep data, as well the the study population.
Results: There has been a steep increase in the number of publications in this research area in the past 10 years. Clustering is the most commonly used method, but other methods are gaining popularity. Apart from classical polysomnography, data from wearable devices, nearables, video, audio, and medical imaging techniques have been used as input to unsupervised machine learning. The broad search allowed us to explore various applications within sleep research ranging from the general population to populations with various sleep disorders.
Conclusion: The review mapped existing research on unsupervised learning in sleep research, identified gaps in the literature, and derived directions for future research.

Key words: Unsupervised Machine Learning, Sleep, Scoping Review

Statement of Significance
Sleep is a transdisciplinary research field. With the rise of unsupervised machine learning and its emergence in sleep research, there is a pressing need to cultivate a mutual understanding across disciplinary boundaries to curate meaningful applications of unsupervised machine learning. This scoping review aims to serve as a foundation to facilitate collaboration across disciplines and ultimately contribute to the elevation of sleep research, by identifying novel ways of applying unsupervised machine learning.

Introduction

Driven by recent advancements in technology, sleep scientists now have the tools to measure and analyze sleep in unprecedented depth and precision. Both supervised and unsupervised machine learning models can take part of the credit for these developments. While supervised learning tends to receive more attention due to its high prediction accuracy, unsupervised machine learning offers a powerful set of techniques, particularly valuable in the medical field, where it can be difficult or expensive to acquire manual labels. LeCun, Bengio, and Hinton predicted in their 2015 review that unsupervised machine learning would gain importance in the long term, as this way of learning from data is similar to the natural learning process of humans and animals [1]. These methods are often overshadowed by supervised machine



learning, even though current research is rapidly turning the gaze toward unsupervised machine learning. Examples of breakthroughs based on unsupervised machine learning in recent years are generative models, which gained a lot of attention for creating realistic images, videos, and music [2]. Other forms of unsupervised learning have been used to train large language models, such as Bert and Chat-GPT, which learn from vast amounts of text data without specific labels, enabling them to generate human-like text [3]. These models have also been applied in the medical context [4]. Based on this overall rising interest in unsupervised machine learning, the question of its contribution to sleep research should be raised. We identified seven existing literature reviews on machine learning in sleep research, and even though they give valuable insights, most of them focus on supervised machine learning models, as can be seen in Table 1.

Table 1. Existing literature reviews about machine learning in sleep research and the number of publications using unsupervised machine learning in each review.

Reference	Review Topic	# of Reviews
Bazoukis et al [5]	OSA	3 of 132
Salari et al. [6]	OSA	3 of 48
Ferreira-Santos et al. [7]	OSA	0 of 63
Gutierrez et al. [8]	Pediatric OSA	0 of 19
Li et al. [9]	Sleep Positions	3 of 27
Sri et al. [10]	Sleep Staging	1 of 19
Alsolai et al. [11]	Sleep Staging	13 of 157

Multiple papers review how machine learning is used to diagnose obstructive sleep apnea (OSA) [5–8]. Li et al. [9] reviewed the automatic recognition of sleep postures using different machine learning approaches. Multiple publications, e.g., [10, 11], provide literature reviews on the most common application of machine learning within sleep research: the automatic scoring of sleep stages. While these reviews give important insights into their respective field, they are mostly focused on the use of supervised machine learning, and no systematic review has been published addressing the use of unsupervised machine learning in sleep research. We aimed to create a holistic view of machine learning in sleep research by covering different kinds of unsupervised machine learning methods and sleep data. By addressing both the methodological breadth and the wide variety of potential use cases within sleep research, this scoping review provides a comprehensive summary of how unsupervised machine learning can be used to advance sleep research.

In the following sections, we will first explain unsupervised machine learning and its significance for sleep research. Then, the most relevant concepts and methods related to unsupervised machine learning will be explained using examples from sleep research. In the method section, we will describe the process of conducting the scoping review, including the search strategy, the inclusion and exclusion criteria, the study selection, and data extraction. Thereafter, we present the results of the review with an overview of all included papers and then analyze the publications based on their different applications within sleep research.

Background

Machine learning is a field of study that allows computers to learn from data. *Learning* in this context means creating an internal model of the data, which can be used to identify patterns or make predictions about new data. There are two main approaches to achieve this: supervised and unsupervised machine learning. Supervised machine learning relies on labeled data, e.g., polysomnography (PSGs) with

manual sleep scoring, for training a model and making predictions. Unsupervised machine learning, on the other hand, uses unlabeled data, e.g., PSGs that have not been scored. Therefore, unsupervised machine learning can infer patterns within data without reference to known or labeled outcomes [12]. The unlabeled training of unsupervised machine learning models does include various methods such as clustering, dimensionality reduction, anomaly detection, and association rule learning [13]. Furthermore, different generative models can be considered as unsupervised machine learning [14]. Summarizing from multiple resources [12, 15] we derive the following definition:

Definition of unsupervised machine learning:
Every machine learning method that does not rely on labeled data.

Another form of machine learning, which is closely related to unsupervised machine learning, is *self-supervised machine learning* [16]. This method trains on input-output pairs similar to supervised learning but generates the labels automatically based on the input data [17]. The last form of machine learning included in this review is semi-supervised learning [18], which can be seen as a mix of supervised and unsupervised learning [19]. This learning approach uses a small amount of labeled data to guide the unsupervised learning process. To cover all the different ways of training a machine learning model without labels, we also included publications that apply self-supervised learning and semi-supervised learning in the review.

The strength of these machine learning methods lies in their independence from manual labels. Especially in sleep research, refraining from using manual labels in sleep staging is desirable for three reasons: (i) manual labels require time-intensive work from a highly skilled professional, (ii) manual labels can have high inter-scorer variability [20–22], and (iii) training on these manual labels will not find new insights but replicate the rules developed by Rechtschaffen and Kales [23] they are based on. Even though these rules are the accepted standard in sleep scoring [24], they are limited by oversimplification as described by Himanen and Hasan [25]. Therefore, it would be desirable to use methods in sleep research, which are not dependent on these labels. Although unsupervised machine learning shows immense potential for sleep research, there is a gap in the literature concerning a review providing a comprehensive overview of the literature published to date outlining the intersection between unsupervised machine learning and sleep research. For this reason, we aimed to review the entire body of literature existing on unsupervised machine learning in sleep research. This scoping review has the following objectives:

1. To provide understanding and guidance in the diverse landscape of machine learning methods
2. To investigate and map the current application of unsupervised machine learning in sleep research both by learning method and data type.
3. To identify potentials for future research based on gaps in the literature and the temporal progression of research trends

In the following sections, we will explain the core methods of unsupervised machine learning. This way, we aim to provide the technological background knowledge to understand the application of unsupervised machine learning in the context of sleep research.

Clustering

The core concept of clustering is identifying groups of elements that are more similar to each other than to the rest of all elements. In the context of sleep research, for example, we could create clusters of similar patients or clusters of similar epochs in sleep staging. In the example of OSA phenotyping, the elements are OSA patients, and the clusters are phenotypes [26]. The important property of clustering, which makes it an unsupervised learning method, is that the labels of these clusters do not need to be known prior to the clustering [27]. The forming of clusters happens only based on the characteristics of each element. In the OSA example, the characteristics of the elements could, e.g., be sleep parameters like the Apnea-Hypopnea Index (AHI) and sleep efficiency. The way these characteristics are used to group similar elements varies between clustering methods [28]. To understand the methods, it is helpful to imagine the elements as points in a coordinate system, where the x and y axis represent two characteristics, as can be seen in Figure 1a).

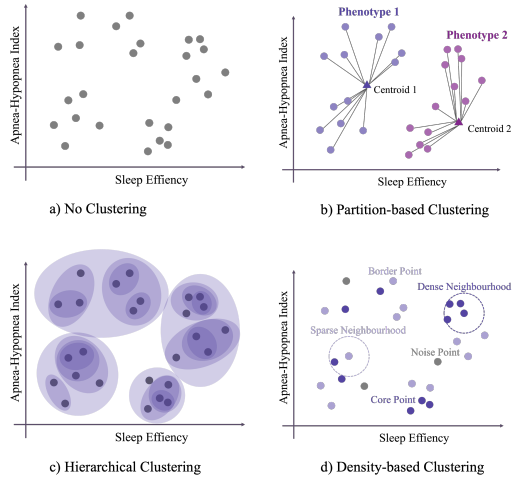


Fig. 1. Examples of Clustering Methods

Partition-based Clustering:
In this method, the data is divided into distinct groups. The most common strategy for finding these groups is called *K-Means*, where the number of groups, which is referred to as *K*, is defined beforehand [29]. Then, K randomly generated data points representing the center of a cluster called *centroids*, are created, and each element is assigned to the centroid with the most similar characteristics as can be seen in Figure 1b). This way, the initial clustering is set, which will then be refined in an optimization process. The similarity between all elements in a cluster and the distance to other centroids is calculated and adjusted until an optimum is found. A very popular variation of K-Means, which is often used in sleep staging, is Fuzzy C Means, which works by similar principles but gives each data point a probability of belonging to each cluster [30].

Hierarchical Clustering:
This method builds a tree-like structure of clusters, where

clusters are merged or split step-by-step [31]. The most common way of building this structure is called *Agglomerative Clustering*, where, at first, each element is treated as a separate cluster. Then, the two most similar clusters are identified and merged. This process can be repeated until only one cluster, including all elements, is left. Figure 1c) visualized hierarchical clusters up to the 4 biggest clusters. It is a useful method when the number of clusters is not known beforehand, as it allows us to explore various grouping levels. In some applications, the tree structure, which is called a *Dendrogram*, reveals important information about the elements and the hierarchy between them.

Density-based Clustering:
Clusters are formed based on the density of neighboring data points [32]. An example of density-based clustering can be seen in Figure 1d). Before the clustering, we need to define: (i) the minimum similarity of two points to be considered in the neighborhood of each other and (ii) the minimum number of points that have to be in a neighborhood to consider it as densely populated. Then each element can be categorized as a *core point*, which fulfills the minimum number of points in their neighborhood, as a *border point*, which has one other element in their neighborhood, or as a *noise point*, which has no other element in their neighborhood. Based on these categories, borders between clusters and points that do not belong to any cluster can be defined. The most common implementation of this method is Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [33]. This clustering method is useful when the data contains noise or outliers.

Dimensionality Reduction

In sleep research, we often deal with complex data types. A PSG, e.g., captures sleep simultaneously using various biosignals, each often with over one hundred measurements per second. As for humans, this complex data can be a challenge for some machine learning methods. The goal of dimensionality reduction is to reduce the complexity of data while still preserving the most important information [34]. Dimensionality reduction can be useful to visualize data, remove noise and minimize storage space and computing time. Furthermore, applying dimensionality reduction to data before using it with other machine learning methods can improve the model performance. Common methods for dimensionality reduction are Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Autoencoders.

Principal Component Analysis:
PCA transforms the original data into a new set of dimensions, which are referred to as the principal components [35]. The new components are combinations of the original features that aim to capture the largest possible variance within the data. The components are sorted by how much variance within the data they capture. This method is used to reduce the dimensionality of data by keeping only a subset of the components.

Independent Component Analysis:
ICA aims to identify and separate independent sources [36]. It is important to mention, that even though the names sound similar, ICA and PCA have different goals and different underlying methods. While PCA aims to combine data into meaningful components, ICA aims to separate data into meaningful components. The key assumption of ICA is the independence of these different sources and that these sources are non-Gaussian. An example of a signal which consists of separate components is the electroencephalography (EEG) signal. It combines the activity arising from different processes in the brain, as well as other processes that may be included in the signal, such as movement or cardiac artifacts. ICA is often used in sleep research to preprocess EEG signals.

Autoencoder:
Autoencoders are a type of neural network that consists of an encoder and a decoder [1]. Figure 2 shows this architecture. On the left side, we can see the input data for the neural network. In this example, the input data are 10-second sequences of a single-channel sleep EEG. Since EEG is typically sampled at a frequency of 200 Hz or higher, each input signal would have a dimensionality of (1,3000). The *encoder* compresses the data into a smaller dimensionality. The data is forwarded through multiple layers of neurons, which learn a representation of the data. When the data has passed all layers of the network and is compressed to the desired size, it is saved in its new low-dimensional representation. In the example in Figure 2 we reduced each EEG sequence to only two values.

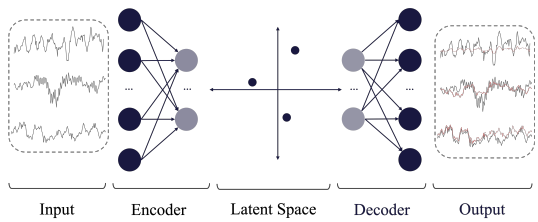


Fig. 2. Architecture of an Autoencoder

We can think of this latent space as a coordinate system, where the two axes represent the two most important features the network identified. These features are abstract but may represent natural features such as amplitude and frequency in this EEG example. Therefore, in the center of the neural network, the bottleneck, the data in our example, will only have a dimensionality of (1,2). The data will then be forwarded to the *decoder*, where it will be layer-by-layer brought back to its original dimensionality. The output of the model, the reconstructed data, will then be compared to the original data. The difference between the original and the reconstructed data is called the reconstruction error and is used as an optimization metric to train the neural network. The data will be passed through the network many times, and in each round, the neurons will be adjusted to reconstruct the data as accurately as possible.

Association Rule Mining

Association rule mining is a machine learning method for discovering relationships between items in a data set based on their co-occurrence [37]. Items that often co-occur are called *frequent item sets* and are used to identify rules. These rules have an if-then structure, i.e., *if* item x occurs, *then* it is very likely that item y will occur as well. In this one-directional relationship, the first item is called the antecedent, and the second item, which is likely to occur as well, is called the consequent. An example would be a dataset with sleep diary

entries including information about sleep quality and sleep hygiene. The items in this data set would be, for example, screen time, naps, or caffeine intake, and they would be co-occurring if they belong to the same night. The association rules derived from this data set could then be, for example, *if* caffeine intake was high, *then* it is likely that sleep quality is low. In this rule example, the antecedent is caffeine intake, and the consequent is sleep quality. These rules are generated based on three metrics which can be calculated for any combination of items:

- **Support:** Measures how often items co-occur, i.e., how many diary entries mention both high caffeine intake and low sleep quality among all diary entries.
- **Confidence:** Measures how often the consequent occurs among all occurrences of the antecedent, i.e., in how many entries low sleep quality is mentioned among all entries that mention high caffeine intake.
- **Lift:** Measures the strength of the association rule by comparing the likelihood of the rule to the random chance of co-occurrence.

The most common method of discovering association rules is the *apriori algorithm* [38]. First, frequent item sets need to be identified, and then rules on the directional relationships between these items can be derived. Beforehand we need to set a minimum support defining how often a set of items needs to appear to be considered a frequent item set. Frequent item sets are found by calculating the support of every possible combination of two items. All item sets that fulfill the minimum support will be saved. Then, gradually, more items will be added to the saved sets, and only the ones that fulfill the minimum support will be saved and further tested. This process continues until no more frequent item sets can be found. Then, the association rules are found by calculating the confidence for all possible association rules within a frequent item set and retaining only rules that meet the minimum confidence. Finally, the lift is used to filter significant association rules from association rules that are only coincidental.

Generative Models

Generative models learn patterns in data to generate new but similar data. There are both supervised and unsupervised generative models. Popular unsupervised generative models are General Adversarial Models (GANs) [39]. This unsupervised machine learning model consists of two components, the *generator* and the *discriminator*. The generator receives real data as an input and learns to generate artificial data. The discriminator receives both real and artificial data and learns how to distinguish between them. Both parts of the model are optimized simultaneously, which allows them to leverage each other. The generator learns to create more and more realistic new data and the discriminator becomes a more and more critical reviewer in each optimization round.

Hidden Markov Model

A Hidden Markov Model (HMM) is a method that uses an observed sequence to predict an unknown sequence [40]. The unknown sequence is assumed to be a *markov chain*, a sequence of states, where the probability of the following state is only based on the current state. An example could be sleep staging, where the unknown sequence is the sleep stages, and the observed sequence is the EEG. HMMs furthermore work with *transition probabilities*. Based on the given data, they calculate the probability of each state to move into any other state. This way, the HMM considers the sequential nature of the data [41].

Unsupervised Anomaly Detection

In anomaly detection, we distinguish between normal and anomalous data. Typically, the majority of data is normal, and the minority of data is anomalous [42]. With anomalous, we refer to data points that are significantly different from the normal. In respiratory data, for example, an anomaly could be coughing with extreme deflections in the signal. However, not all anomalies have to be extreme values. A data point can also be anomalous, with values in a normal range but in an unusual context. For example, OSA patients might switch to paradoxical breathing during an airway obstruction. The abdomen signal itself is not unusual, but in the context of the thorax signal, it is anomalous. There are different unsupervised anomaly detection methods, many of them relying on the clustering and dimensionality reduction concepts discussed in the previous section [43]. The density-based clustering method DBSCAN can, for example, be used where points in low-density areas are considered anomalous. Furthermore, the autoencoder can be used where the data points with a high reconstruction error are considered anomalous [44].

Method

Scoping Review

Scoping review is a type of systematic review used to map and analyze a field of research [45]. We did a qualitative assessment of the research output in an emerging field by describing its main characteristics and publication trends. This protocol was elaborated according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) and its extensions to protocols and scoping reviews [46]. The protocol is available at Open Science Frameworks [47].

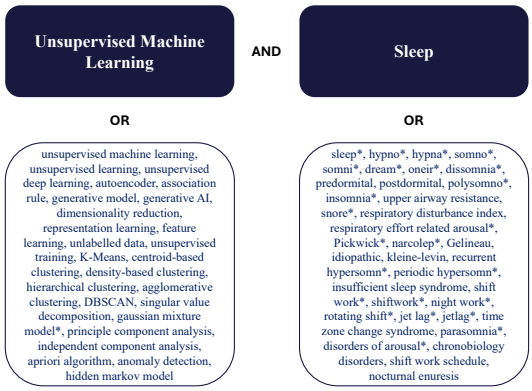


Fig. 3. Search Terms for Unsupervised Machine Learning and Sleep

Search Strategy

The literature search was performed in four databases: PubMed, Web of Science, Scopus and ACM Digital Library. The search strings were initially developed for PubMed and

then adapted to the other databases' syntax and search engines. The search strategy was composed by combining two components: unsupervised machine learning (methods) and sleep (application domain). To increase the search sensitivity, no restrictions on the type of sleep study, outcomes, or population were set. A combination of MeSH terms and free terms were used for each of the search components. The search string for the sleep domain consisted of the search filter proposed by Pires et al. [48]. Gray literature and secondary data sources were not screened. An overview of the included search strings and the general search query logic is shown in Figure 3.

Inclusion and Exclusion Criteria

The eligibility analysis was based on the following inclusion criteria:

- **Language:** Only papers published in English were considered eligible. Papers in any other languages were excluded.
- **Type of Paper:** Reviews, philosophical or conceptual research, editorials, opinion papers, letters to the editor, and non-peer-reviewed papers, e.g., posters and book chapters, were excluded.
- **Population:** Only papers studying humans were considered eligible. Animal studies were excluded.
- **Sleep:** Only papers primarily related to sleep were considered eligible. A paper was considered to be sleep-related if the study population, intervention, exposition factor, or main outcome were related to sleep.
- **Unsupervised machine learning:** Only papers presenting an application of an unsupervised machine learning method on sleep-related data were considered eligible. The abstract should either mention the term unsupervised, unlabeled data, or unsupervised training or mention an unsupervised machine learning method.

Study Selection

Our screening process was done according to the PRISMA guidelines. The number of publications at each step of the process can be seen in Figure 4. The records retrieved from the literature search in the four databases were imported into Rayyan [49], where de-duplication and eligibility analyses were performed. Duplicate records with a similarity above 95% were excluded automatically, and all remaining identified duplicates were checked manually by the first author. From initially 7043 records, 5004 possible duplicates were identified. 3083 of these duplicate records were removed. Hence the screening was performed on the remaining 3960 records. All non-duplicated papers were evaluated in a two-step process. The abstracts were screened by the first author and three independent reviewers. Before the start of the screening, each reviewer screened 100 records to ensure the screening instructions were clear enough to provide a sufficient agreement. All publications that were not in agreement in the pilot screening phase were discussed, and the screening instructions were updated accordingly. This pilot screening resulted in a Cohen's Kappa of 0.54 with reviewer 1, 0.69 with reviewer 2, and 0.62 with reviewer 3. Then, the full set of publications was screened.

All of the abstracts were screened by the first author. Furthermore, reviewer 1 screened 50% of the records, while reviewers 2 and 3 screened 25% of the records each. This way, each record was screened by two independent reviewers to limit individual bias. The screening resulted in an agreement of 0.67 across all three reviewers, which is considered a substantial agreement. The conflicts were resolved by discussing each conflict between the first author and the respective reviewer. All of the 856 included abstracts in this screening round used unsupervised machine learning methods, although they did not focus on the method; instead, they focused on the sleep-related contribution. Therefore, another abstract screening round conducted by the first author, selecting only publications where unsupervised machine learning was the main contribution, resulted in 440 papers, which were then included in the full-text retrieval.

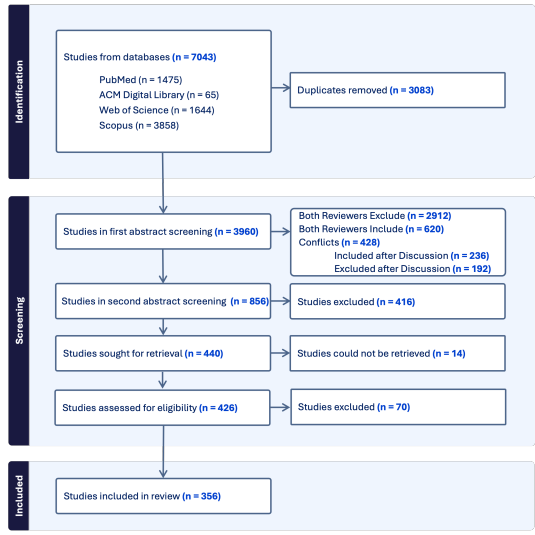


Fig. 4. Flow Chart According to the PRISMA Guidelines

Data Extraction and Analysis

The data extraction was done in multiple steps. In the first step, we coded all 856 included publications based on their abstract. We coded them by the unsupervised machine learning method and the type of sleep data. This step allowed us to get an overview of methods and data types across all related publications. We furthermore coded the abstracts by focusing on the purpose of unsupervised machine learning. This includes whether unsupervised machine learning was the main contribution of the paper and whether it was a novel usage of unsupervised machine learning or a known methodology. We continued the in-depth data extraction process only with publications that focused on unsupervised learning and had a methodological contribution. This way, we got a broad overview of unsupervised machine learning in sleep and were then able to further narrow down the most relevant applications for the full-text review.

In the second step, the full text of the 440 included publications were reviewed. The majority of these publications were retrieved using the automatic retrieval of open-access papers with Endnote. The remaining 156 papers, which could not be retrieved this way, were manually downloaded. 14 publications, which were not open access and were not provided

by the authors within two weeks after contacting them, were excluded from the sample. Once the final sample of eligible studies was reached, data extraction was performed. In 67 of the studies, only the full-text review revealed that they were not eligible in terms of language, method, or data type. For each of the 356 included publications, the information listed below was extracted.

- **Study Description:** papers metadata (first author, publication year, source title, and full reference string).
- **Source Country:** The country of the first affiliation of the first author.
- **Unsupervised Machine Learning:** The method of unsupervised machine learning was the main focus of the paper.
- **Role of Unsupervised Machine Learning:** The purpose or role of using unsupervised machine learning in the publication.
- **Data Type:** The type of sleep data used by the unsupervised machine learning method.
- **Clinical Outcome:** The application of the unsupervised machine learning method within sleep research.
- **Data Set**: The data set used, population characteristics, and overall number of individuals in the study.
- **Results:** The evaluation metric and performance of the machine learning model. Furthermore, whether the author concluded their research as successful or not.
- **Outlook:** The limitations or future research directions the paper mentions.

There were no mandatory or conditional items, and missing items were filled as "non-available/not applicable". We used both predefined categorical labels as well as free text fields for extracting the information. This way, both a quantitative analysis and a qualitative analysis of the reviewed papers were possible.

Results

The first part of the results is dedicated to providing a comprehensive overview of all the papers included in this review, categorizing them by publication year, employed method, and type of data utilized. The second part of the results section is structured according to the various application areas within sleep research.

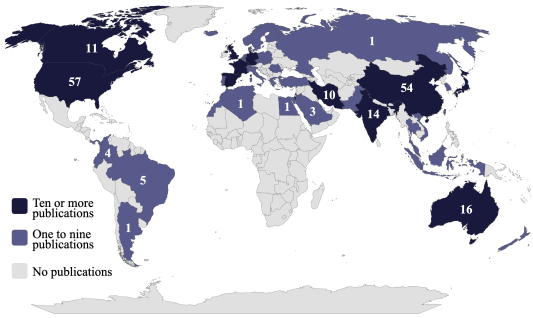


Fig. 5. Publication Countries Based on First Author Affiliation

Countries

We extracted the country of the first author's affiliation for each reviewed publication with the goal of identifying the most involved locations in this research area. Figure 5 shows a map colored by the number of publications from the respective country. Most publications stem from the United States, China, and Australia. All blue countries have at least one publication, and all dark blue countries have more than 10 publications.

Timeline

The research on unsupervised machine learning in sleep research gained increasing popularity in recent years. Figure 6 shows a steep ascent in publications since 2016. This can be explained by the generally increased research interest in machine learning across all application areas and disciplines over the last decades [50]. More than 40 publications on unsupervised machine learning in sleep research were published in 2023 alone. However, the timeline also shows that the theoretical foundations for applying these methods date back to the 1970's.

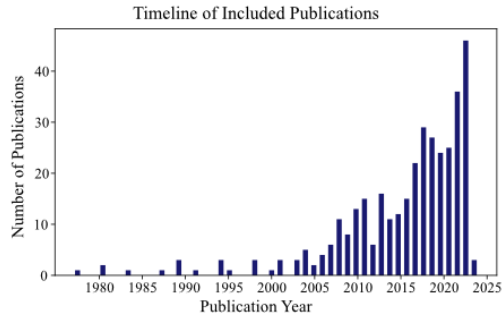


Fig. 6. Number of Publications by Year

Multiple of these early publications used fuzzy clustering for sleep staging [51–56]. Other research used X2-based clustering [57], hierarchical clustering [58], self-organizing maps [59] and ISODATA clustering, an iterative self-organizing clustering method [60] to classify sleep stages. Apart from sleep staging, unsupervised machine leaning has been used for modeling the process of drifting into sleep [61] and extracting features from sleep data to classificy spindles [62, 63] in the 1990s. Early applications of K-Means in sleep research have been removing artifacts [64] and detecting micro arousals [65] in sleep EEG. Clustering is clearly the most widely used unsupervised learning method in sleep research. When zooming in on the publications from the last 20 years, a trend toward other unsupervised machine learning methods becomes visible. The number of publications using dimensionality reduction methods has been rising since 2016. In the past three years, we can see that other methods are gaining popularity, which include methods such as unsupervised domain adaptation or contrastive learning.

Data Types

The scoping review showed that the papers across all sleep applications varied strongly in the way data was collected. As a first step, we mapped the different data types into the categories (i) wearables and nearables, (ii) physiological data,

and (iii) meta data and other data. We logically divided the data types into measurement devices that can be used for long-time monitoring in a home setting, such as wearables, nearables, audio, video, and other types of sensors on the one hand, as seen in Figure 8, and medical devices, which are typically used for one night studies and require the assistance of a professional, seen in Figure 9 on the other hand. We furthermore created one category that includes different forms of metadata and other data types in Figure 10. We counted the number of publications which use each data type in their work for each category and subcategory. Publications using multiple data types are counted multiple times in this visualization.

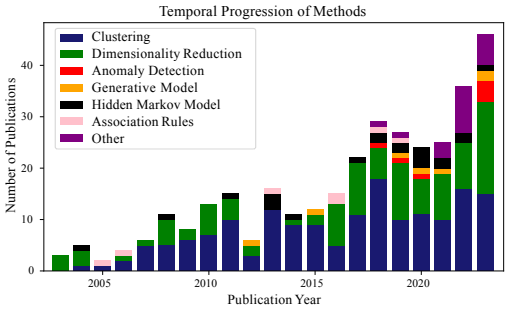


Fig. 7. Methods used in Publications between 2003 and 2023

Wearables and Nearables

Figure 8 gives an overview of the different types of wearable and nearable devices to measure sleep. In the included papers, smartwatches have been used for sleep staging [66–68], detect respiratory events [69] and analyze sleep patterns [70]. Different brands of smartwatches have been used, including Apple watch [66], FitBit [67, 71, 72] and Withings ScanWatch [73]. Actigraphy has been used for sleep-wake monitoring [74, 75], OSA screening [76] and identifying anomalous sleep patterns [77]. Furthermore, an Oura Ring has been used to collect longitudinal sleep data [78]. There have been 7 publications using different types of mattress sensors [79–82] to assess sleep. There were publications using sensors in the four corners of the bed [83], in the pillow [84] or woven into the bed sheet [85].

Video has been used to identify sleep postures [86], monitor breathing [87–89] during sleep and monitor vigilance when driving a car [90, 91]. While most papers used simple cameras, some research experimented with different types of cameras, including infrared camera [88, 92] and 3D camera [93]. There are two major ways that audio data was used: (i) monitoring during sleep and (ii) extracting information about sleep from speech during wake. When monitoring sleep, a microphone is typically placed on the body or close to the bed and sounds like coughs [94] or snoring [95, 96] were identified. In most cases, the microphone is the one native to smartphones, but it may also be used as part of other nearable and wearable devices. Audio was used to detect respiratory events [97–99] or predict the age of a sleeping person [100]. When extracting information from speech - and therefore in wake state - the publications used the audio from interviews or experimental settings. Many papers succeeded in predicting OSA severity from speech recordings during wake [101–103].

There are various types of sensors to collect information on sleep. Shahid et al. [104] made use of recent developments in the smart home industry. They drew data from the motion sensors of elderly individuals living in smart homes. They showed that these sensors, primarily designed to control the light, could also be used to gather information about sleep in the form of bed and rise times, as well as awakenings during the night. A different publication used a mattress sensor for sleep monitoring and then communicated with the smart home devices to adjust light and temperature for optimal sleeping conditions [105]. Gu et al. [106] showed that Wifi sensors, typically used to provide internet access, can be reused for sleep monitoring [106]. Based on the channel state information and received signal strength, they can detect movement in sleep with an accuracy of 98.2%. Non-commercial products such as Doppler radar have been used to decompose the respiration signal of two people in one bed [107]. Other forms of data include ultra-wideband radar, which was used to identify anomalous sleep patterns [108], a piezoelectric sensor attached to the neck, which tracks snoring vibrations [109] or an infrared sensor, which is used to classify movements in bed [110]. In drowsiness detection, an air quality sensor has been used to predict changes in driver alertness [111].

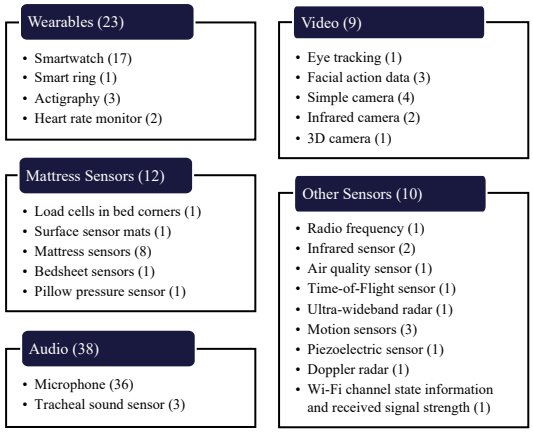


Fig. 8. Overview of Data from Wearables and Nearables

Clinical Physiological Signals

The most common way to capture sleep is PSG, which allows measuring several physiological signals during sleep. We categorize signals by the part of the body they measure, such as the brain, the heart, and the breathing. The most common way to measure the brain in sleep research is EEG. A total of 191 of the 356 publications used EEG as an input to unsupervised machine learning. There are different forms of EEG offering varying levels of invasiveness. Wireless EEG [112] is a minimally invasive measurement for monitoring brainwaves with fewer attachments. In contrast, intracranial EEG is a more invasive method, where electrodes are placed directly inside the skull. This technique is primarily used in clinical or research settings, e.g., for conditions like epilepsy, where precise localization of brain signals is crucial [113, 114]. Functional Magnetic resonance imaging (fMRI) offers more detailed representations of brain structure and activity compared to EEG, providing high-resolution images that can capture subtle brain changes

during sleep. The two included publications used fMRI to remove ballistocardiographic artifacts [115] and modeling brain states from sleep to wake [116].

In the study of cardiovascular function during sleep, different methods offer varying levels of precision. There were 40 publications using cardiovascular measurements as an input for unsupervised learning. Wearable devices, such as smartwatches, commonly measure heart rate through optical sensors, providing a general but less detailed view of heart activity [70, 117–119]. Electrocardiography (ECG) electrodes on the chest offer a more precise measurement of heart function during sleep. The ECG signal has been used in unsupervised machine learning for sleep-wake monitoring [120], drowsiness detection [121] and sleep staging [122]. Blood pressure measurements can also be relevant in research on morning surge [123] or analyze the relationship between blood pressure and OSA [76].

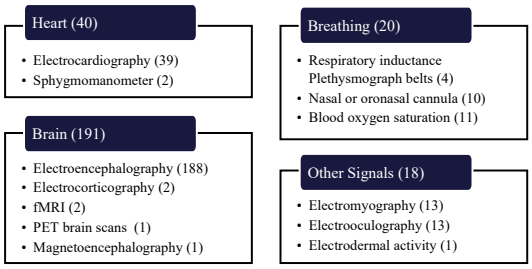


Fig. 9. Overview of Physiological Data

20 reviewed publications included respiration, which can be captured using a nasal or oronasal cannula [124–127]. Another way to capture respiration is through respiratory movements, using Respiratory Inductance Plethysmograph (RIP) belts around the thorax and abdomen [128–130]. Reimer et al. [131] uses a skin conductive electrode belt that captures both respiratory rate and heart rate. Another physiological process is blood oxygen saturation, which is influenced by breathing, which is why many OSA-related publications use pulse oximeters to detect respiratory events [132–135]. 13 publications used electromyography (EMG) and electrooculography (EOG), which are essential for detecting Rapid Eye Movement (REM) sleep. Electrodermal activity (EDA) is used to measure skin conductance as an indicator of sweating, offering insights into the autonomic nervous system's activity during sleep. Daley et al. [136] use EDA to analyze the body in prolonged wakefulness.

Metadata and Other Data

The term metadata describes data that provides information about other data. In this category, we summarized data sources that provide information about a person and their sleep but are not direct, continuous measurements of sleep. Figure 10 shows different types of metadata. This includes, for example, hypnograms or sleep parameters derived from a PSG recording. Self-reported sleep data such as sleep questionnaires and sleep diaries also count into this metadata category. There are publications that include general health information in their analysis, such as demographics [137–139], anatomical information [140], lifestyle information [127, 141], co-morbidity [142], or medical history [143]. Examples of publications using only metadata of sleep staging are Mirth et al. [144] or Jouan et al. [145], who analyze the sleep stage scoring provided by multiple sleep experts instead of analyzing the raw data itself.

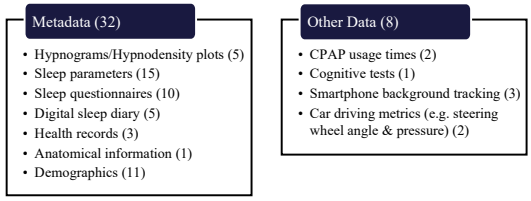


Fig. 10. Overview of Meta Data and Other Data

Some data types do not fit in any of these categories. Baddam et al. [146] and Kang et al. [147] tracked the adherence to Continuous Positive Airway Pressure (CPAP) usage of OSA patients. Based on current usage hours, they categorized OSA patients and predicted future adherence. Boyraz et al. [148] extract driving metrics such as steering wheel angle or break behavior to identify drowsy drivers. Massar et al. [78] use the background tracking information from smartphones, such as tapping or usage hours, to analyze sleep patterns. Rošťáková and Rosipal [149] used objective cognitive tests to identify sleep types.

Unsupervised Machine Learning Methods

The most commonly used unsupervised methods are clustering and dimensionality reduction. To cover the entire field of unsupervised machine learning, we also included other methods, such as unsupervised anomaly detection, association rules, generative models, and HMMs. The Sankey diagram in Figure 11 shows how the number of publications of the different method types and data types is distributed. Additionally, we can see the flow, showing how many publications from each method category are used with which data type.

Fig. 11. Sankey Diagram Showing the Flow between Method Types and Data Types (HMM = Hidden Markov Model, PSG = Polysomnography)

We identified 181 publications using clustering. The most common clustering methods used are K-means clustering, Gaussian Mixture Models (GMMs), Fuzzy C-means clustering, and hierarchical clustering. We reviewed 113 publications that focused on dimensionality reduction. It is important

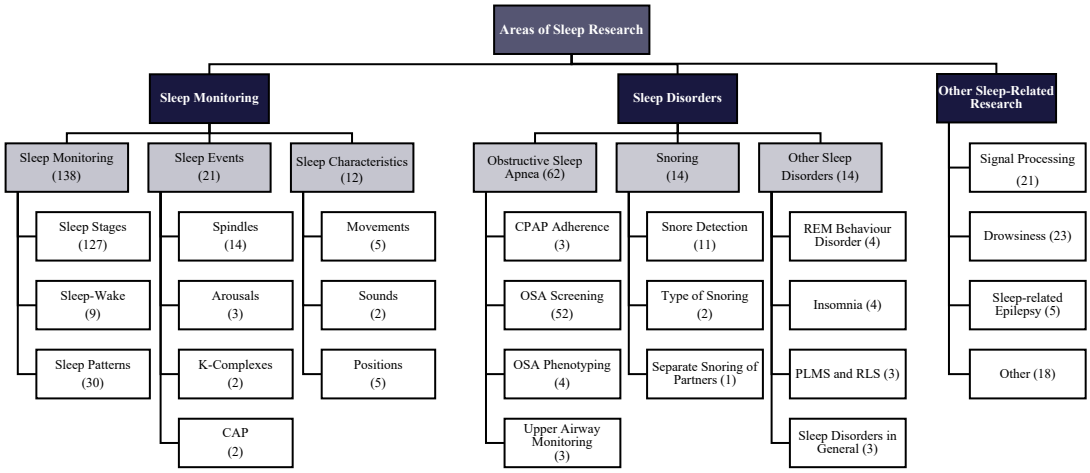


Fig. 12. Categorization of Publications in Different Areas of Sleep Research

to mention here, that dimensionality reduction techniques are already a commonly used method for preprocessing in medical research. The most common forms of dimensionality reduction in sleep research were PCA, ICA, Autoencoders, Singular Value Decomposition, Self-Organizing Maps, and Deep Belief Networks. They were most often used for classification, preprocessing, feature extraction [150], pre-training of a supervised model [151, 152] and data compression [128, 153].

There were 22 publications using HMMs with different purposes, such as modeling the process of drifting into sleep [61], modeling sleep transitions [154], and cycling alternating pattern (CAP) analysis [155]. The review includes 8 publications that used unsupervised anomaly detection [44, 77, 79, 108, 111, 156–158] in the context of sleep research. Some of them aimed to identify risks based on sleep patterns, such as risk during pregnancy [79] or behavioral risk for the elderly [158]. Other publications used anomaly detection methods for drowsiness detection [111, 157] or identifying mouth breathing during the night [44]. There are 7 publications included in the review that create association rules based on sleep data [71, 72, 159–163]. Most of them were used for explorative analysis of clinical data sets. Álvarez, Félix, and Cariñena applied this method to the scoring of respiratory events to identify breathing patterns during sleep. We identified 7 publications using generative models in sleep research [95, 164–169]. Most of these publications aimed to generate artificial sleep data as a method to improve the classification performance for spindle detection [164, 167], snore detection [95] and sleep staging [168]. Other research aimed to explore the sleep data through these generative models [169] or create art [165]. There are 11 publications with other methods which could not be included in any of the aforementioned categories. One of these methods is unsupervised domain adaptation [170–178], where a model trained on a labeled source domain can adapt to a different but related target domain that lacks labeled data. Other methods are a competition convolutional neural network [179], a Hierarchical Multi-Agent System [180], and a Bayesian switch-point model [181].

While the majority of studies employed unsupervised machine learning, 14 studies applied semi-supervised machine learning. Additionally, 16 publications focused on self-supervised methods. Most of these publications used contrastive learning, which compares positive pairs of similar data (e.g., different representations of the same sleep cycle) against negative pairs (e.g., different individuals' sleep data), helping the model learn distinguishing features. For example, Xiao et al. [182] used contrastive learning to predict sleep stages.

Research on Sleep Monitoring

An overview of different areas of sleep research and the number of publications in each category can be seen in Figure 12. The term sleep monitoring in this context aims to describe any type of measuring physical attributes during sleep or tracking sleep patterns over a period of time.

Sleep Staging

Sleep staging is the most common classification task in sleep research. 129 of the 356 included publications focus on sleep staging. We review the machine learning methods, data types and data sets, population size (labelled as $\#$) and characteristics, and evaluation in these publications in Appendix 1. Different unsupervised machine learning methods have been applied for sleep staging. The most common methods are various variations of K-Means clustering with 25 publications and different variations of autoencoders with 19 publications. The most common roles of unsupervised machine learning in sleep staging besides classification are domain adaptation and feature extraction. Some publications use unsupervised machine learning in unique ways, such as weighting features [183] and clustering the train set [184, 185]. Tian et al. [186] use unsupervised machine learning to refine the the accuracy of a supervised model. They use clustering to classify only the epochs that the supervised classifier is not sure about.

Most publications use EEG for sleep staging. Others use simplified measurement channels, for example, integrated into wearables. The most common other signals are ECG [168, 187], Photoplethysmography [66, 188], accelerometer [67, 131, 189]

and video [92, 93]. Vanbuis et al. [130] predict the sleep stages of participants with sleep-disordered breathing using only respiratory signals and the heart rate and achieve an accuracy of 0.79. Reviewing the data sets used for training and testing the automated sleep staging models showed that 49 of 129 publications on sleep staging rely on the publicly available Sleep-EDF data set. This data set is available through physionet.org and consists of PSG recordings of healthy adults. An extended version with healthy and sleep-disordered adults is available as well. The recordings took place between 1989 and 1994. Other common public data sets are The Montreal Archive of Sleep Studies (MASS) data set, St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (UCD), ISRUC data set [190], and the Sleep Heart Health Study (SHHS) data set [191]. Even though the goal of using public data sets is often comparability, the reviewed publications often use subsets of the public data sets or combine multiple public data sets.

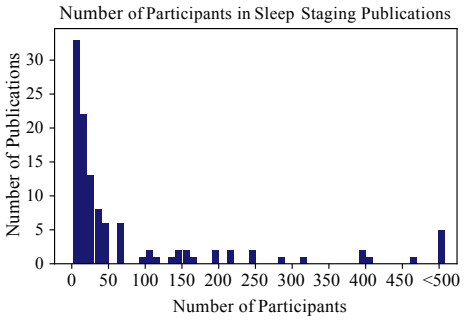


Fig. 13. Number of Participants in Sleep Staging Publications

Specifically, 69 of 112 sleep staging papers, which provide information about their population, train and validate their models on a healthy population only. Most of the publications use the data of adults or don't specify the age range. There are 7 publications doing sleep staging with the data of elderly citiziens, 6 publications with the data of children and 7 with the data of new borns. The size of the data sets used in sleep staging publications can be seen in Figure 13. Most publications train and validate their models on the data of one to ten participants. There are five publications which use data sets with 500 participants or more.

Sleep-Wake Monitoring

Sleep-Wake monitoring can be used to calculate sleep parameters, such as the sleep duration, sleep efficiency, sleep onset latency and wake after sleep onset. It can be used to track sleep longitudinally and monitor for example regularity of sleep and wake times. An overview of all publications using unsupervised machine learning for sleep-wake detection can be found in Table 2. While the previous chapter showed that automatic sleep staging is usually based on EEG, the publications on sleep-wake monitoring mostly use wearables or nearables. This allows a longitudinal tracking of sleep wake rhythms. Some of the included publications use the longitudinal sleep-wake patterns for an early diagnosis of neurological disorders or to detect other anomalies.

Reference	Method	Data Type	Population	#	Duration	Metric	Value
El-Khadiri et al., 2018 [181]	Bayesian Switch Point Model	Motion sensors	Healthy adults	1	219 nights	Accuracy	0.94
Geng et al., 2022 [120]	Shapelets and K-Means	ECG	Healthy adults, suspected SDB	30	1 night	Accuracy	0.78 - 0.88
Rai et al., 2015 [192]	Fuzzy C-Means	EEG	Healthy adults	8	1 night	Accuracy	0.95
Liu et al., 2020 [119]	HMM	Smartwatch heart rate and accelerometer	Elderly	14	3 months	Agreement	0.87
Muns et al., 2017 [74]	K-Means	Actigraphy	Healthy adults	10	80 days	-	-
Fung et al., 2023 [118]	K-Means	Smartwatch heart rate	Healthy adults	1	1 night	Accuracy	0.97
Shahid et al., 2022 [193]	K-Means	Motion sensors	Elderly	6	3 years	Silhouette Score	0.4
Subramanian and Coleman, 2022 [194]	K-Means and HMM	Smartwatch accelerometer	Healthy adults	7	1 day	Accuracy	0.93
El-Manzalawy et al., 2017 [75]	K-Means, Fuzzy C-Means and GMM	Actigraphy	Insomnia patients and healthy adults	37	3-11 nights	Accuracy	0.85

Table 2. Publications on Sleep-Wake Monitoring

Sleep Patterns

With the technological enhancements in wearable and nearable technologies, we can observe a growing interest in longitudinal analysis of sleep patterns across multiple nights. Long-term monitoring can identify irregular sleep behaviors or deviations from normal sleep patterns, which may be indicative of underlying health issues [77, 108]. The duration of sleep monitoring also varies, from one-night studies to longitudinal studies spanning up to a year. An overview of all publications on sleep patterns can be found in Table 3. Other publications use unsupervised learning to analyse [73] and predict [141] sleep quality. Clustering has been applied to identify group of individuals with similar sleep patterns based on demographics or sleep parameters. The sample sizes vary significantly across the literature, ranging from small cohorts of 7 participants [70] to large-scale studies with up to 2,579 individuals [139].

Sleep Events

Apart from sleep stages, there are other events visible in EEG, such as sleep spindles, K-complexes, and arousal. Table 4 provides an overview of all publications using unsupervised machine learning methods to detect sleep spindles. All of the reviewed papers used EEG signals to detect spindles. Piza et al. [209] used kernelized K-Means clustering to sample the training set before using a supervised model for classification. Loza et al. [167] uses hierarchical clustering to create a structure of patterns from vector spaces of different dimensions. Chen et al.[210] classify spindles using K-Means with an Accuracy of 0.927. There are two publications that use unsupervised machine learning to classify K-Complexes. Ranjan et al. [211] uses fuzzy C Means clustering, and Zacharaki et al. [212] use spectral clustering to detect K-Complexes with an accuracy of 0.912 and 0.84. Arousals are another event in sleep EEG and can be defined as a sudden shift in EEG frequency that indicates a brief disruption in sleep. They are difficult to identify with supervised machine learning since the agreement of manual scoring is low [21]. Identifying arousals with unsupervised machine learning was first approached by Pacheco and Vaz

in 1998 by using K-Means clustering [65]. In recent years, Safont et. al have attempted to classify arousals with Sequential Independent Analysis Mixture Models [213, 214]. They tested their model on three participants with OSA and reported a classification accuracy of 0.8. There are two publications on cyclic alternating pattern, which are both by Mendonça et al. [155, 215]. Their first publication is based on a single-lead ECG signal. They use a Deep Stacked Autoencoder to predict sleep quality based on the Cycling Alternating Pattern [215]. The second publication extracts features from EEG signals and uses a HMM, a GMM, and Self-Organizing Maps to estimate the Cycling Alternating Pattern [155].

Sleep Characteristics

Sleep Movements

Detecting movements during sleep has been approached with various measurement devices. Table 5 provides an overview of the different methods and data types used to classify movement

Reference	Method	Role	Data Type	Population	#	Duration
Caroppo et al., 2018 [108]	Incremental Clustering	Detect anomalous sleep patterns	Accelerometer, Time-of-Flight sensor and ultra-wideband radar	Healthy adults	18	3-5 months
Liang et al., 2016 [71]	Association Rules	Discover factors relevant to sleep quality	Smartwatch	Healthy adults	2	30 days
Huijben et al., 2022 [195]	Contrastive Predictive Coding and Self-organizing Maps	Exploratory analysis of sleep patterns	EEG, EOG and EMG	Healthy adults	96	1 night
Massar et al., 2021 [78]	K-Means	Explore sleep behaviour	Oura ring, smartphone background tracking app and digital sleep diary	Healthy adults	198	8 weeks
Laxminarayan et al., 2005 [160]	Apriori Algorithm	Derive association rules	Sleep questionnaire, demographics and sleep parameters	Suspected sleep disorders	81	1 night
Kim et al., 2019 [161]	Apriori Algorithm	Derive association rules	Sleep questionnaire, demographics and sleep parameters	Suspected sleep disorders	81	1 night
Liang et al., 2016 [72]	Apriori Algorithm	Derive association rules	Smartwatch	Healthy adults	5	180 days
Laxminarayan et al., 2006 [162]	Apriori Algorithm	Derive association rules	Sleep questionnaire, demographics and sleep parameters	Suspected sleep disorders	81	1 night
Gasmi et al., 2023 [77]	Mean Shift Algorithm and Autoencoder	Identify anomalous nights	Actigraphy	Elderly	1	1 year
Jansen et al., 1987 [57]	X2-based Clustering	Sleep patterns	EEG and EOG	OSA patients and healthy adults	25	1-2 nights
Li et al., 2023 [196]	K-Means	Predict sleep health indicators	Mobile application (audio and usage activity)	Healthy adults	1	4 years
Alabdan et al., 2023 [197]	Stacked Sparse Autoencoder	Predict sleep quality	Mobile application	Healthy adults	1	1.5 years
Khumngoen et al., 2023 [198]	Principal Component Analysis and K-Means	Predict sleep quality	-	-	-	7 weeks
Zhang et al., 2017 [199]	Transfer Learning using Deep Autoencoder	Sleep quality prediction	EEG and smartwatch	Healthy adults	10	1 night
Hong et al., 2017 [141]	Deep Belief Network	Sleep quality prediction	IoT devices, demographics, daily lifestyle reports	Healthy adults	333	2 weeks
Wu et al., 2017 [200]	Self-organising Maps, Hierarchical Clustering and Hidden Markov Model	Sleep quality prediction	Microphone and questionnaire	Healthy adults	36	1 night
Park et al., 2023 [201]	K-Means and Collaborative Filtering	Sleep recommendations	Smart bed	-	-	-
Wang et al., 2013 [202]	Expectation Maximization Clustering	Identify sleep patterns	EEG, demographics and sleep questionnaire	Suspected sleep disorders	244	1 night
Lee et al., 2022 [139]	Autoencoder and K-Means	Identify sleep patterns	Demographics and sleep questionnaires	Soldiers with sleep disturbances	2579	-
Usher et al., 2012 [203]	[]Expectation Maximization Clustering	Identify sleep patterns	Hypnogram	Healthy adults	244	1 night
Wang et al., 2017 [204]	Combined Dynamical Modeling Clustering	Identify sleep patterns	Hypnogram	Healthy adults	244	1 night
Bajkowski et al., 2023 [205]	Fuzzy Evidence Accumulation Clustering	Identifying sleep types	Respiration, activity and heart rate	Elderly	19	One year on average
Usami, 2014 [137]	Constrained K-Means	Identifying sleep types	Sleep parameters and demographics	Junior highschool children	100	3 years (annualy)
Khasawneh et al., 2010 [206]	Gaussian Mixture Model	Identifying sleep types	Sleep parameters, demographics and health factors	Healthy adults	244	1 night
Mirth et al., 2023 [144]	K-Means, K Modes and Principal Component Analysis	Identifying sleep types	Hypnogram and hypnodensity plot	Healthy adults	98	1 night
Rošťáková et al., 2019 [149]	K Medoits with Dynamic Time Warping	Identifying sleep types	EEG, self-reported sleep, vigilance test and blood pressure	Healthy adults	146	2 nights
Khasawneh et al., 2011 [207]	Gaussian Mixture Model	Identifying sleep types	Sleep Parameter	Suspected sleep disorders	244	1 night
Alfeo et al., 2018 [70]	Fuzzy Clustering	Identifying sleep types	Heart Rate, Accelerometer and Sleep Quality Reporting	Healthy adults	7	20 nights
Biedebach et al., 2023 [73]	K-Means	Identifying sleep types	Smartwatch and Digtial Sleep Diary	Healthy adults	45	3 months
Wallace et al., 2018 [208]	Mixture Model based on Multivariate Skew Normal Distribution	Identifying sleep types	PSG, Smartwatch and Self-reported Sleep Data	Older adults with and without insomnia	216	1 week

Table 3. Publications on Sleep Patterns

Reference	Method	Population	#	Duration	Metric	Value
Patti, Penzel and Cvetkovic, 2015 [216]	Gaussian Mixture Model	Healthy adults	6	1 night	Sensitivity	0.57
Patti, Chaparro-Vargas and Cvetkovic, 2014 [127]	Gaussian Mixture Model	Healthy adults	6	1 night	Sensitivity	0.75
He et al., 2022 [128]	Variational Switching State-Space Model	Healthy adults	1	2 nights	-	-
Chen et al., 2021 [129]	K-Means	Sleep disordered adults	6	1 night	Accuracy	0.93
Rosipal et al., 1998 [62]	ICA	-	1	7 min	-	-
O'Reilly et al., 2015 [220]	Hierarchical clustering	Healthy adults	9	1 night	Sensitivity	0.85
Loza et al., 2021 [164]	Deep Neural Dynamic Bayesian Network	Healthy adults	55	30 min	Accuracy	0.42
Piza et al., 2017 [209]	Kernelized K-Means	Healthy adults	27	1 night	Sensitivity	0.86
Caspary et al., 1994 [63]	Singular Value Decomposition	-	-	-	-	-
Chen et al., 2023 [221]	K-Means	Sleep disordered adults	20	1 night	Accuracy	0.90
Ventouras et al., 2010 [222]	ICA	Healthy adults	1	1 night	-	-
Patti et al., 2018 [223]	Multivariate Gaussian mixture model	Healthy adults	25	1 night	Sensitivity	0.74
Ventouras et al., 2008 [224]	ICA	Healthy adults	1	1 night	-	-
Loza et al., 2019 [167]	Hierarchical Clustering	-	8	30 min	Sensitivity	0.68

Table 4. Publications on Sleep Spindles

during sleep. Wi-Fi [106], radio frequency [225], infrared sensors [110], and video [226] have been used to monitor sleep movements from a distance. These different publications defined different classes of movement. While two publications only differentiate between movement and no movement, or roll-over or no-rollover, others identify more specific movements. Adami et al. [83] used load cells in the corners of the bed to classify a movement as a major posture shift, a smaller movement in the upper body, or leg movement. Bagci Nguyen and Ozturk [110] differentiate between respiration, twitches, limb movements, and tossing and turning. All of the mentioned publications use clustering either in the form of K-Means or GMMs.

Reference	Method	Data Type	Movement	#	Duration	Acc.
Gu et al., 2019 [225]	Gaussian Mixture Model	Radio frequency-based monitoring	General Movement	11	Short time	0.97
Adami et al., 2011 [83]	Gaussian Mixture Model	Load cells in bed corners	Major posture shifts, smaller movements in upper body or legs	15	Short time	0.85
Bagci et al., 2023 [110]	K-Means	Infrared sensor	Respiration, twitches and limb movements, tossing and turning	1	Short time	0.88
Heinrich et al., 2013 [226]	K-Means	Video	Turning, stretching legs, moving arms and head	1	Short time	0.67
Gu et al., 2020 [106]	Gaussian Mixture Model	Wi-Fi	Roll-overs	7	1 hour	0.98

Table 5. Publications on Sleep Movements

Sleep Sounds

Sounds can reveal information about sleep. Sounds related to sleep-disordered breathing will be treated in the sections on OSA and snoring. In this section, we review publications that aim to detect sounds during sleep in general. Barata et al. [94] aim to detect coughs during sleep. They use a GMM to cluster audio segments. They even take one step further, to differentiate between coughs of partners sleeping in one bed. Their work was tested on 94 participants with asthma over 28 nights each and resulted in a Matthews correlation coefficient of 0.92. Another work on sound classification during sleep is by Wu et al. [227]. They classify sounds during the night as tooth grinding, snoring, movement, or environmental noise. They use self-organizing maps and test the performance of this unsupervised method on 7 healthy adults for one night of audio recording. Their model classifies the event with a pairwise F Measure of 0.581.

Sleep Postures

During sleep, we switch between different sleeping postures or positions. Evaluating sleeping positions might be relevant in the evaluation of sleep disorders, such as postural obstructive sleep apnea (i.e., sleep apneas occurring only in the supine position) [228] or restless sleep disorder (a sleep-related movement disorder characterized by frequent and large movements during sleep) [229]. Research has aimed to automatically identify these postures with unsupervised machine learning, as can be seen in Table 6. The papers either differentiate between supine, i.e. lying on the back, left side and right side [81, 82, 86] or additionally classify prone, i.e., lying on the stomach or on the side [85]. The overarching goal of the sleep posture classification can be relevant to monitoring the health status of the elderly [86] or assessing the risk of pressure ulceration in bed-bound patients [81].

Reference	Method	Role	Data Type	Positions	#	Metric	Value
Baran Pouyan et al., 2015 [82]	Fuzzy C-Means	Classification	Surface sensor mats	Supine, left, right	10	Accuracy	0.93
Ostadabbas et al., 2014 [81]	GMM with EM Algorithm	Classification	Mattress sensor	Supine, left, right	9	Accuracy	0.98
Bhatlawande et al., 2022 [86]	K-Means	Feature Extraction	Images	Supine, left, right	109	F1 Score	0.92
Hsiao et al., 2018 [85]	Fuzzy C-Means	Feature Extraction	Bedsheet sensor, infrared sensor	Supine, prone, left lateral, right lateral, left side, right side	-	Accuracy	0.88

Table 6. Publications on Sleep Postures

Research on Sleep Disorders

There are multiple publications that analyze sleep disorders in general [230, 231]. For example, Bruce [232] uses K-Means to cluster EEG sequences by participants with insomnia, nocturnal frontal lobe epilepsy, periodic leg movements, and REM behavior disorder. This research showed that the clustering was useful in identifying oscillatory patterns in the EEG of these sleep disorders and neurological disorders.

Obstructive Sleep Apnea

The review shows that most publications on Obstructive Sleep Apnea (OSA) focus on automatically scoring respiratory events, including apneas and hypopneas, during the night. There are multiple publications using Gaussian Mixture Models to predict OSA from Speech [101–103, 256–261]. The choice of data type varies significantly across studies. For instance, 18 publications rely on audio recordings to capture respiratory patterns, 6 use SpO2 data, and 17 utilize ECG signals. A full overview

Reference	Method	Role	Data Type	Population	#	Duration	Metric	Value
Haidar et al., 2019 [128]	Autoencoder	Classify Respiratory events	RIP belts and cannula	-	2056	Subset of a night	Accuracy	0.81
Moeynoi et al., 2017 [233]	Canonical Correlation Analysis	Classify Respiratory events	ECG	OSA patients	25	1 night	Accuracy	0.90
Biedebach et al., 2024 [44]	Convolutional Autoencoder	Classify Respiratory events	Different respiration signals	Children with and without OSA	20	1 night	F1 Score	0.51
Hu et al., 2023 [234]	Convolutional Autoencoder	Classify Respiratory events	ECG	OSA patients and healthy adults	95	1 night	Accuracy	0.90
Almarshad et al., 2023 [135]	Convolutional Autoencoder and Transformer Neural Network	Classify Respiratory events	SPO2	Middle aged adults	30	1 night	Accuracy	0.80
Mostafa et al., 2017 [132]	Deep Belief Network	Classify Respiratory events	SpO2	OSA patients	33	1 night	Accuracy	0.85 - 0.98
Li et al., 2020 [134]	Dirichlet Process Mixture Model	Classify Respiratory events	SpO2	-	33	1 night	Accuracy	0.85 - 0.97
Le et al., 2013 [117]	Dirichlet Process-based Mixture Gaussian Process Model	Classify Respiratory events	ECG and wearables	OSA patients and healthy adults	26	1-2 nights	Accuracy	0.77 - 0.88
Feng et al., 2021 [235]	Frequential stacked sparse auto-encoder and Hidden Markov Model	Classify Respiratory events	ECG	OSA patients	32	2 - 4 nights	Accuracy	0.851
Ravelo-García et al., 2004 [236]	Gaussian Mixture Model	Classify Respiratory events	ECG and SpO2	OSA patients and healthy adults	66	1 night	Accuracy	1
Goldshtein et al., 2011 [99]	Gaussian Mixture Model	Classify Respiratory events	Speech audio	OSA patients and healthy adults	83	1 night	Specificity, Sensitivity	0.79, 0.83
Elmoaqet et al., 2020 [125]	Gaussian Mixture Modeling	Classify Respiratory events	Oronasal airflow	OSA patients	96	1 night	Accuracy	0.80
Ben-Israel et al., 2010 [97]	Gaussian Mixture Modeling	Classify Respiratory events	Sleep audio	Healthy adults	60	1 night	Sensitivity	0.92
Sim et al., 2022 [237]	Greedy Pre-pruned Tree-based Clustering	Classify Respiratory events	-	-	-	-	Accuracy	0.92 - 0.99
Al-Ani et al., 2008 [238]	Hidden Markov Model	Classify Respiratory events	ECG	OSA patients	70	1 night	Accuracy	0.7
Novák et al., 2004 [239]	Hidden Markov Model	Classify Respiratory events	EEG	-	-	-	-	-
Feng et al., 2019 [240]	Hidden Markov Model and Sparse Autoencoder	Classify Respiratory events	ECG	-	70	1 night	Accuracy	0.85
Ostadieh et al., 2020 [241]	Hybrid Radial Basis Function using K-Means	Classify Respiratory events	ECG	OSA patients and healthy adults	70	1 night	Accuracy	0.96
Zhao et al., 2011 [98]	K-Means	Classify Respiratory events	Sleep audio	Snorers and OSA patients	42	1 night	Sensitivity, Specitivity	0.90, 0.92
Boudaoud et al., 2005 [242]	K-Means	Classify Respiratory events	ECG	OSA patients	5	30 min	Sensitivity	0.84
Boppana et al., 2019 [243]	K-Means combined with KNN	Classify Respiratory events	ECG	-	-	1 night	Accuracy	0.97
Marcos et al., 2008 [133]	K-Means with RBF	Classify Respiratory events	SpO2	Suspected OSA	187	1 night	Accuracy	0.86
Alvarez et al., 2007 [244]	K-Means, Hierarchical Clustering and Fuzzy C-Means	Classify Respiratory events	Pulse oximetry	Suspected OSA	74	1 night	Accuracy	0.91
Robertson et al., 2007 [245]	Principal Component Analysis, Empirical Mode Decomposition	Classify Respiratory events	Airflow	OSA patients	3	1 night	Sensitivity	0.81
Ostadieh et al., 2020 [246]	Hybrid RBF network with K-Means	Classify Respiratory events	ECG	OSA patients and healthy adults	70	1 night	Accuracy	0.96
Kumar Tyagi et al., 2023 [247]	Restricted Boltzmann Machine in Deep Belief Networks	Classify Respiratory events	Single lead ECG	OSA patients	70	1 night	Accuracy	0.89
Kumar et al., 2023 [248]	Self-supervised representation learning	Classify Respiratory events	ECG	OSA patients	95	1 night	Accuracy	0.85
Takao et al., 2019 [69]	Stacked Autoencoder	Classify Respiratory events	Mattress sensor	Healthy adults	5	2.5 minutes	Accuracy	0.90
Li et al., 2018 [249]	Stacked Sparse Autoencoder and Hidden Markov Model	Classify Respiratory events	Single lead ECG	OSA patients	70	1 night	Accuracy	0.847
Zubair et al., 2023 [250]	Sub-pattern-based PCA	Classify Respiratory events	ECG	OSA patients	70	1 night	Accuracy	0.87 - 1
Sepúlveda-Cano et al., 2011 [251]	Time-adapted Principal Component Analysis	Classify Respiratory events	PPG	Children with suspected SDB	21	1 night	Accuracy	0.83
Sharma et al., 2020 [252]	Variational Mode Decomposition	Classify Respiratory events	ECG	OSA patients	70	1 night	Accuracy	0.88
Alshaer et al., 2009 [253]	K-Means	Creating segments of Breath	Sleep audio	OSA patients	-	-	-	-
Joergensen et al., 2021 [124]	Agglomerative Hierarchical Clustering	Identify patterns in breathing	Airflow, SpO2 and heart rate	Healthy adults	10	1 night	Accuracy	0.64
Álvarez et al., 2013 [154]	Apriori Simple Temporal Problem Miner	Identify patterns in breathing	Scoring of respiratory events	OSA patients	50	1 night	-	-
Boudaoud et al., 2007 [254]	K-Means	Identify patterns in breathing	ECG		7	2 minutes	Sensitivity, Specitivity	0.81, 0.84
Holm et al., 2023 [129]	Variational Autoencoder and K-Means	Identify patterns in breathing	Airflow and RIP belts	OSA patients and snorers	100	1 night	-	-
Temrat et al., 2018 [255]	Fuzzy C-Means	Predict OSA from Audio	Tracheal sound	OSA patients	49	1 night	Accuracy	0.88
Ren et al., 2020 [76]	Multivariate Dirichlet Process Mixture	Analyze OSA and Blood Pressure	Blood pressure monitoring cuff and actigraphy	Children with and without OSA	97	1 night	-	-
Wong et al., 2023 [142]	Principal Component Analysis	Predict OSA from Metadata	Sleep parameters, demographics, comorbidities	Cancer patients	249	1 night	F1 Score	0.91
Zigel, Tarasiuk and Goldshtein, 2008 [256]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	26	1 night	Accuracy	0.92
Blanco et al., 2009 [257]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	26	Short time measurement	Error Rate	0.078
Blanco et al., 2011 [258]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	-	Short time measurement	Relative Reduction in EER	0.25
Pozo et al., 2009 [259]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	80	Short time measurement	Accuracy	0.81
Elisha et al., 2011 [260]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	Suspected OSA	92	1 minute	Sensitivity, Specitivity	0.92, 0.92
Blanco et al., 2013 [101]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	80	Short time measurement	Accuracy	0.89
Fernández et al., 2009 [103]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	80	Short time measurement	Accuracy	0.81
Gómez-García et al., 2013 [102]	Gaussian Mixture Model Variations	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	520	Short time measurement	Accuracy	0.65
Fernández et al., 2010 [261]	Gaussian Mixture Model	Predict OSA from Speech	Speech audio	OSA patients and healthy adults	80	Short time measurement	Accuracy	0.81

Table 7. Publications on OSA

Reference	Method	Role	Data Type	Population	#	Duration	Metric	Value
Schmitt et al., 2016 [270]	Deep autoencoder and HMM-GMM	Snore detection	Microphone	Healthy adults	44	1 night	F1 score	0.95
Wongsirichot et al., 2016 [269]	Degenerate unmixing estimation echnique	Separate snoring from partners	Microphone	Adults with suspected sleep disorders	110	Short time measurement	Mean source to interference ratio	12.83
Mordoh and Zigel, 2021 [268]	Fuzzy 2-Means	Classify type of snoring	Microphone	Adults with suspected OSA or snoring	15	1 night	-	-
Azarbarzin and Moussavi, 2011 [271]	Gaussian Mixture Model	Snore detection	Directional microphone	Adults with OSA	33	1 night	Detection Rate	0.97
Dafna et al., 2011 [272]	GMM with EM Algorithm	Snore detection	Smartphone microphone	Healthy adults	6	1 night	Accuracy	0.91-0.8
Romero et al., 2019 [267]	Hierarchical Clustering	Snore detection	Piezoelectric sensor	Healthy adults	156	2 nights	F1 Score	0.93
Yadollahi et al., 2009 [273]	ICA	Snore detection	Microphone	-	1	Short time measurement	-	-
Vrins et al., 2004 [274]	K-Means	Snore detection	Smartphone microphone	Healthy adults	5	1 night	Accuracy	0.75
Goh et al., 2018 [109]	K-Means	Snore detection	Microphone	Healthy adults	15	1 night	-	-
Beeton et al., 2007 [96]	K-Means "random++" variant	Classify type of snoring	Nasopharyngoscope and headset microphone	Adults with OSA and snoring	24	1 night	Unweighted Average Recall	0.80
Zhang et al., 2020 [95]	K-HarmoniC-Means Clustering	Snore detection	Microphone	Adults with OSA	1	1 night	Accuracy	0.96
Azarbarzin et al., 2010 [275]	PCA and Fuzzy C-Means	Snore detection	Ambient microphone and tracheal microphone	Adults with suspected OSA	30	1 night	Accuracy	0.99
Bublitz et al., 2017 [276]	PCA and K-Means	Snore detection	Microphone	Adults with suspected OSA	20	1 night	Accuracy	0.956
Ma et al., 2015 [277]	Semi-supervised Conditional GAN	Snore detection	Microphone	-	-	-	Unweighted Average Recall	0.52

Table 8. Publications on Snoring

of publications using unsupervised machine learning to detect respiratory events is shown in Table 7.

Other publications use clustering to identify phenotypes of OSA. They use metadata such as sleep parameters, health records, lifestyle information, and self-reported sleep information as a basis for the clustering [138, 143, 262]. Some publications monitor the upper airway using audio [263, 264] or a static-charge-sensitive bed and a nasal cannula [126]. There are publications that specifically aim to identify the side of upper airway collapse [68, 265]. Three publications analyze [146, 266] the CPAP usage times of OSA patients and predict their future adherence [147] .

Snoring

We identified 14 publications using unsupervised machine learning to classify snoring. An overview of the used machine learning methods, data sets, and performance metrics can be found in Table 8. All of these publications use audio data from microphones except Romero et al. [267], who use a piezoelectric sensor. Most of the research aims to detect snoring events, while others more specifically try to classify the type of snoring [96, 268] or separate the snoring of partners [269].

Insomnia

There are four publications using unsupervised machine learning focusing on insomnida. Park et al. [278] cluster people with insomnia based on sleep patterns collected with a smartwatch. They aim to detect different types of insomnia and run their model on 6 weeks of longitudinal data by 42 adults with insomnia. There are two publications aiming to detect insomnia based on EEG recordings [174, 279]. Frederic et al. [280] use ICA for EEG signal preprocessing, specifically for people with insomnia.

REM Sleep Behavior Disorder

There are multiple applications of unsupervised machine learning methods in the context of REM sleep behavior disorder (RBD). Tripathi and Rajendra [281] use PET brain scans to predict whether a person with idiopathic RBD is likely to develop Parkinson's or Lewy Bodies dementia. Koch et al. [282] aim to classify people with RBD using EEG-based sleep staging. There are two publications that analyze the sleep of people with RBD based on PPG and accelerometer data from wearables [188] and EMG [283].

Restless Legs Syndrome and Periodic Limb Movement Disorder

Restless Legs Syndrome (RLS) is characterized by an uncomfortable urge to move the legs [284], often worsening at night, while PLMD involves repetitive, involuntary leg movements in the daytime and during sleep [285]. There are two publications that use clustering to identify movements based on mattress sensor data [286] and EMG [287]. Fairly et al. [288] use PCA to extract features from EMG data, to detect phasic EMG activity.

Research on Other Sleep-related Topics

Drowsiness

Vigilance analysis aims to detect sleepiness or microsleep during wake time, often in the context of driving. A full overview of publications using unsupervised machine learning to assess drowsiness can be found in Table 9. A substantial number of studies in this field concentrate on physiological signals. Some employ EEG to monitor brain activity, while others utilize ECG to analyze heart rate variability. Furthermore, researchers have attempted to derive levels of sleepiness from speech patterns [289, 290]. More recent advancements involve the real-time collection of data within the vehicle itself, including the tracking of facial expressions [90, 91], eye movements [136], and steering behavior [148], as well as the monitoring of environmental factors such as air quality [111].

Signal Processing

Unsupervised learning has been used to extract and decompose signals. ICA and PCA have been used to extract respiratory

rate from video [87, 89]. Additionally, both respiratory rate and heart rate have been extracted from infrared video [88], piezoelectric sensors [80], and pillow pressure sensors [84]. Signal decomposition can also be applied to separate the respiration of two people in one bed [107]. Poreé et al. [291] applied ICA to extract EEG, EMG, and EOG from a simplified EEG set-up. ICA is not only useful for decomposition but also for data compression. Crainiceanu et al. [292] developed a dimensionality reduction method they call population value decomposition, which they use to compress EEG signals. They work with the data of the SHHS data set, including 3201 recordings. This compression is desirable for storing the data and for subsequent data analysis.

Back in 1989, Lima, Leitao, and Paiva [64] used K-Means to remove artifacts from EEG. Today, most publications use ICA [293–295]. More specifically, ICA has been used to clean EEG data by removing muscle artifacts [296, 297], cardiac artifacts [298–300] and eye movement artifacts [301]. Somervail et al. [302] use Artifact Subspace Reconstruction to clean the EEG signal. These methods can also be used to create new data for signals with faulty [303] or missing [304] data.

Sleep-related Epilepsy

Four of the reviewed papers focus on epilepsy patients. Lee et al. [115] use ICA to remove ballistocardiographic and ocular artifacts from EEG and model the hemodynamic response functions. The overall goal of the research is to use simultaneous EEG and fMRI recording to analyze brain areas generating interictal epileptic discharge spikes during sleep onset. The other publications aim to detect epileptic seizures during sleep based on EEG [305] and use intracranial EEG to detect sleep stages [113] and sleep-wake states [306].

Other

Several studies have employed sleep measurements to assess risk in specific population groups, such as pregnant women [79] and elderly [158]. Wang et al. [79] utilized sleep data from smartwatches and sleep diaries collected throughout pregnancy to predict high-risk pregnancies. They detected anomalous sleep patterns, reporting a Spearman correlation of 1.125 between sleep metrics and high-risk pregnancy outcomes.

Several studies have utilized unsupervised learning methods to conduct more general exploratory analyses of sleep independent of specific population groups or sleep disorders. Notably, Abeysinghe and Cui [163] explored large-scale clinical datasets, combining data from multiple cohorts, both with and without sleep disorders, amounting to a total of 24,515 clinical records. By applying association rule mining, they identified patterns within the data, some of which align with existing medical literature and others that may present novel hypotheses for future investigation.

Other publications have applied unsupervised learning techniques to gain deeper insights into physiological processes during sleep, such as cardiac activity [307] or the transition process from wakefulness to sleep [61]. Houldin et al. [116] employed ICA to compare resting-state networks in wakefulness and sleep using fMRI data. Another interesting work using EEG signals called *the Dream Catcher experiment* aimed to detect markers of dreaming consciousness with evidence accumulation clustering but did not succeed [308].

Shifting focus from physiological signals, several publications have analyzed sleep on a meta-level through the examination of hypnograms. Jouan et al. [145] applied a multinomial mixture model to analyze sleep stage decisions made by multiple experts, allowing them to estimate the uncertainty in manual scoring for each epoch, highlighting 'grey areas' in sleep staging. Similarly, Bentrup [309] utilized single linkage clustering to identify grey areas in the predictions from automated sleep staging systems. This approach was proposed as a quality assurance method to enhance the reliability of automated sleep staging outcomes. Alvarez and Ruiz [154] developed an HMM based on hypnograms to model the transitions between different sleep stages. Using the hypnograms of 875 participants, they calculated the transition probabilities between sleep stages, providing a probabilistic framework to better understand the dynamics of sleep stage progression.

Generative machine learning methods have been applied in sleep research to improve the representation of EEG signals [310] or generate artificial sleep EEG [169]. Kumi et al. [166] applied a Gaussian copula model to generate synthetic sleep data that mimics longitudinal smartwatch measurements, capturing realistic sleep patterns. Shahid et al. [140] uses PCA to generate interactive visualizations of respiratory signals. Fernandes et al. [165] took a more conceptual approach, using the KANTS clustering algorithm to create art from sleep EEG signals, demonstrating a novel intersection between sleep research and creative data representation.

Discussion

Contribution of Unsupervised Machine Learning

The review showed that unsupervised machine learning has a significant contribution to sleep research. The 356 publications on unsupervised machine learning in the field of sleep research cover a wide range of sleep-related research areas. This showed that sleep research can gain a lot more from unsupervised machine learning than only sleep staging. The current state and aim of unsupervised machine learning is, in most applications, to improve and simplify the work of the human expert but not fully replace them. The review showed the most common ways of simplifying the process of sleep staging, for example, with clustering algorithms to group similar patterns in sleep data automatically, reducing the need for manual annotations. Additionally, dimensionality reduction techniques were employed to condense high-dimensional data into interpretable formats, facilitating quicker insights without significant loss of information. Unsupervised learning can be used to improve the accuracy of supervised machine learning methods, for example with feature extraction or clustering the training set beforehand.

Limitations in Existing Research

Bridging the gap between sleep expertise and machine learning expertise is necessary to create meaningful applications. The expertise of a sleep researcher is needed to fully understand the data to ensure the application has a meaningful contribution for them. The expertise of the machine learning engineer is needed to decide which methods can be considered and how to implement them. We aimed to bring both parties on the same page and start a discourse about unsupervised machine learning in sleep research.

Public Data Sets

The review revealed a clear trend of many publications relying on the same datasets, with the Sleep-EDF dataset being the

Reference	Method	Data Type	#	Duration	Metric	Value
Shi and Lu, 2008 [311]	Dynamic Clustering	EEG	17	Daytime experiment	-	-
Shi et al., 2007 [312]	Extended Graph Factorization Clustering	EEG	16	Daytime experiment	-	-
Staroniewicz, 2021 [289]	Gaussian Mixture Model	Audio of speech	6	Overnight	Equal error rate	0.03 - 0.11
Rajini et al., 2018 [313]	K-Means	EEG	-	-	Accuracy	1
Yin et al., 2011 [314]	K-Means	EEG	1	Daytime experiment	-	-
Gurudath et al., 2014 [315]	K-Means	EEG	12	Overnight	-	-
Rezaee et al., 2013 [90]	K-Means	Facial expressions	4	Daytime experiment	Accuracy	0.93
Fujiwara et al., 2019 [157]	Multivariate Statistical Process Control using PCA	ECG and EEG	34	Daytime experiment	Sensitivity	0.92
Li et al., 2008 [316]	Probabilistic Principal Component Analysis	EEG	10	Daytime experiment	Accuracy	0.96
Fujiwara et al., 2023 [121]	Self-attention Autoencoder	ECG	20	Daytime experiment	Sensitivity	0.88
Sommer et al., 2001 [317]	Self-organising Map Network	EEG	11	Daytime experiment	None	-
Noori et al., 2016 [318]	Self-organising Map Network	EEG	7	Daytime experiment	Accuracy	0.77
Wali et al., 2013 [112]	Subtractive Fuzzy Clustering	Wireless EEG	50	Daytime experiment	Accuracy	0.84
Chung and Kim 2020 [111]	VAE and skip-GAN	Air quality sensor	95	Longitudinal tracking	-	-
Ayyagari et al., 2021 [153]	PCA	EEG	8	Daytime experiment	AUC	0.91
Boyraz et al., 2008 [148]	Fuzzy Subtractive Clustering	Video and car metrics	30	Daytime experiment	Accuracy	0.89
Dutta, Kour and Taran, 2020 [319]	Clustering Variational Mode Decomposition	EEG	16	Overnight	Accuracy	0.97
Schwarz et al., 2023 [91]	HMM, PCA, K-Means, Hierarchical Clustering	Facial expressions and car metrics	40	Daytime experiment	AUC	0.85 - 0.87
Amiriparian et al., 2020 [290]	Recurrent Autoencoder	Audio of speech	915	Daytime experiment	Spearman's Correlation Coefficients	0.367
Maftukhaturrizqoh et al., 2019 [320]	K-Means as Part of a Radial Basis Function Neural Network	ECG	14	Daytime experiment	Accuracy	0.82
Daley et al., 2022 [136]	Gaussian Mixture Model	Eye and face tracking, ECG and EDA	20	Overnight	-	-
Hsu et al., 2017 [321]	ICA	EEG	10	Daytime experiment	AUC	0.75
Leong and Mandic, 2008 [322]	Noisy Component Extraction Algorithm	EEG and EOG	-	-	-	-

Table 9. Publications on Drowsiness

most commonly used [323]. While the Sleep-EDF dataset has been widely used to train and test machine learning models, it is important to critically assess its limitations. Collected in the 1980s, this dataset consists predominantly of healthy, young Caucasian individuals, which raises concerns about its representativeness and applicability to broader, more diverse populations. Especially in clinical applications, where classification performance can have direct consequences on the diagnosis, it is important to consider a diverse population in the training set to create a machine learning model that is equally reliable for patients of all ages, genders, and races [324]. Therefore, although it is used as a benchmarking standard for comparing sleep staging performance, this reliance on a single, outdated dataset limits the generalizability of findings. Additionally, the focus on incremental performance gains, often in the form of small percentage improvements in sleep staging accuracy, suggests that sleep staging may already be a solved problem from a machine learning perspective. These marginal improvements lack significant clinical relevance, as they may not be meaningful for real-world sleep health outcomes.

Population Characteristics

The same diversity issue arises for people with sleep disorders. The review showed that the majority of sleep staging models are validated primarily on healthy participants, despite the fact that individuals with sleep disorders are the population most in need of accurate and reliable sleep staging. More than half of the publications on automatic sleep staging rely on data from healthy participants only, limiting their clinical applicability to the general population. This focus on healthy populations may be due to the availability of clean, easily accessible datasets and the inherent challenges in working with disordered sleep data, which can be more variable and difficult to model. However, this practice leaves a critical gap in sleep research, as models that perform well on healthy individuals may not generalize effectively to those with conditions like insomnia, sleep apnea, or other disorders. For future advancements in sleep staging to have a clinical impact, models must be validated on diverse populations, including those with sleep disorders, to ensure their effectiveness in real-world clinical settings.

Clinical Usefulness of Evaluation

In order to overcome these limitations, we are in need of a public data set that covers a broad population range. This data set should be complemented by an open-access and peer-reviewed data descriptor that clearly describes the data set and the characteristics of the population as well as guiding the evaluation procedure to ensure comparability between publications. Standards for a common evaluation procedure should be designed in a way that shows that the application of unsupervised machine learning makes sense in a clinical context. It should be given that publications aiming to contribute to sleep monitoring, should also test their methods on a whole night sleep recording. Still, multiple publications tested their applications only on short-time recordings during the day. Surprisingly, many applications evaluated their model only on the data of a a single subject. Another common flaw is training the models on 80% of the recording of all participants and then testing the models on the remaining 20% of their recording. This form of evaluation includes data from each participant in the training, which does not reflect the generalizability to completely new participants. Therefore, we suggest a test set including participants of different ages, genders, and pathologies for a comparable and generalizable evaluation.

Limitations of the Review

One limitation of the review is the lack of comparability between studies. However, we intentionally sacrificed this

comparability to widen the scope of included studies. Therefore, our review provides a broad analysis of how unsupervised machine learning has been used in sleep research but does not allow any conclusion regarding their accuracy, efficacy, or practical implementation. A traditional review with meta-analysis, would be feasible if we had focused on one type of study. However, we aimed to create an overview of all different forms of research publications within this field. For this reason, we did not analyze any effects and did not do a bias assessment. In any case, the exploratory nature of our review may serve as an initial step for future meta-analyses of unsupervised machine learning in sleep research, both by our groups and by others.

New Pathways for Sleep Research

By mapping out the entire body of literature on unsupervised learning in sleep research, gaps in sleep research areas became visible. Comparing the number of publications by the sleep disorders they aim to analyze, diagnose, or treat showed a high imbalance. 80 publications were working on topics related to sleep-disordered breathing, including OSA and snoring. Only 19 papers in total considered other sleep disorders and related neurological conditions, including sleep-related epilepsy, insomnia, RBD, PLMS, and RLS. Even though sleep apnea is a highly prevalent and serious sleep disorder, this imbalance does not seem proportional. Furthermore, other sleep disorders such as narcolepsy, circadian rhythm disorders, sleep terrors, somnambulism, or bruxism have not been approached with unsupervised machine learning at all. This may stem from varying knowledge about these disorders, diagnosis ways, and treatment options. For this reason, some of these disorders might have lower demands or no demand for technological support in either of these steps, leading to an imbalance in research. The lack of machine learning-based research on these disorders, may also be lower availability of sleep data from affected individuals. Lastly, this imbalance might also stem from a bias towards disorders with greater commercial potential, such as sleep apnea, which is widely recognized and has a clear market for diagnostic tools and treatments like CPAP machines. The lack of attention to a broader range of sleep disorders points to a need for more balanced research efforts that address a wider spectrum of sleep conditions, particularly those that remain under-represented despite their prevalence and impact on sleep health. We suggest filling these gaps by exploring applications of unsupervised machine learning targeting specifically underrepresented sleep disorders.

Generative machine learning models, association rules, and unsupervised domain adaptation are methods with little research thus far and an increasing number of publications in recent years. Generative machine learning models experienced large technological advancements, which led to an increased interest in these methods in the general population. We showed that only 7 publications used generative models in sleep research so far. These publications were, for example, used to visualize [129] and generate artificial sleep data [169]. There were only 7 applications of association rules, of which 6 were used for explorative analysis of clinical data sets. It would be an interesting research direction to apply them to different data types, similar to how it has been done with respiratory event scoring [159]. In other medical fields, this method has been used to detect risk factors contributing to a specific condition [325, 326]. Another promising method is the unsupervised domain adaptation, which allows us to apply a machine learning model to sleep data we have no or little access to and train the model on a similar data set. This method has, for example, been used for combining MRI data from different centers [327]. We suggest exploring these methods further, which could ultimately reveal new pathways for sleep research.

Conclusion

This scoping review illustrated the diversity of research related to the use of unsupervised machine learning for analyzing sleep and sleep disorders. The various forms of data collection efforts and machine learning methods identified in this scoping review showed that unsupervised machine learning is already an important and embedded part of modern sleep research. Furthermore, our findings show the potential for novel applications in the future by outlining particular pathways that could be of significance for sleep. When analyzing the chronological rise in the number of publications that utilize unsupervised machine learning in sleep research, we can see that the interest is growing rapidly, especially in less-known unsupervised learning methods. This clearly illustrates that unsupervised machine learning research is on the rise, and although its uptake has not yet reached the same heights as supervised machine learning, the tables are turning. This confirms LeCun, Bengio, and Hintons [1, p.7] prediction that unsupervised machine learning may gain greater impact and importance in the long term. Their words, "We discover the structure of the world by observing it, not by being told the name of every object," apply to sleep research as well since unsupervised machine learning allows for the direct discovery of sleep by observing it.

Disclosure statement

Financial Disclosure

This project received funding from the European Unions Horizon 2020 research and innovation program under grant agreement no. 965417. Gabriel Natan Pires is a shareholder at SleepUp©, founder of P&P Metanálises, and receives funding from the Associação Fundo de Incentivo à Pesquisa (AFIP), Brazil. Marie-Ange Stefanos receives funding from Withings France SA. The other authors have indicated no financial conflicts of interest.

Non-financial Disclosure

None.

Data Availability

All relevant data are contained within the article and in the supplementary material. Further inquiries can be directed to the corresponding author/s.

References

1. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.
2. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

3. Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.
4. Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200, 2023.
5. George Bazoukis, Sandeep Chandra Bollepalli, Cheuk To Chung, Xinmu Li, Gary Tse, Bethany L Bartley, Salma Batool-Anwar, Stuart F Quan, and Antonis A Armoundas. Application of artificial intelligence in the diagnosis of sleep apnea. *Journal of Clinical Sleep Medicine*, 19(7): 1337–1363, 2023.
6. Nader Salari, Amin Hosseinian-Far, Masoud Mohammadi, Hooman Ghasemi, Habibolah Khazaie, Alireza Daneshkhah, and Arash Ahmadi. Detection of sleep apnea using machine learning algorithms based on ecg signals: A comprehensive systematic review. *Expert Systems with Applications*, 187:115950, 2022.
7. Daniela Ferreira-Santos, Pedro Amorim, Tiago Silva Martins, Matilde Monteiro-Soares, and Pedro Pereira Rodrigues. Enabling early obstructive sleep apnea diagnosis with machine learning: systematic review. *Journal of Medical Internet Research*, 24(9):e39452, 2022.
8. Gonzalo C Gutiérrez-Tobal, Daniel Álvarez, Leila Kheirandish-Gozal, Félix Del Campo, David Gozal, and Roberto Hornero. Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis. *Pediatric Pulmonology*, 57(8): 1931–1943, 2022.
9. Xianglin Li, Yanfeng Gong, Xiaoyun Jin, and Peng Shang. Sleep posture recognition based on machine learning: A systematic review. *Pervasive and Mobile Computing*, 90: 101752, 2023.
10. Tellakula Ramya Sri, Jahnavi Madala, Sai Lokesh Duddukuru, Rupasri Reddipalli, Phani Kumar Polasi, et al. A systematic review on deep learning models for sleep stage classification. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1505–1511. IEEE, 2022.
11. Hadeel Alsolai, Shahnawaz Qureshi, Syed Muhammad Zeeshan Iqbal, Sirirut Vanichayobon, Lawrence Edward Henesey, Craig Lindley, and Seppo Karrila. A systematic review of literature on automated sleep scoring. *IEEE Access*, 10:79419–79443, 2022.
12. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
13. Samreen Naeem, Aqib Ali, Sania Anam, and Muhammad Munawar Ahmed. An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*, 2023.
14. Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti. A survey of unsupervised generative models for exploratory data analysis and representation learning. *Acm computing surveys (csur)*, 54(5):1–40, 2021.
15. Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.
16. Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, 30(4):2761–2775, 2023.
17. Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.
18. Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.
19. Shahid Tufail, Hugo Riggs, Mohd Tariq, and Arif I Sarwat. Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*, 12(8):1789, 2023.
20. Sami Nikkonen, Pranavan Somaskandhan, Henri Korkalainen, Samu Kainulainen, Philip I Terrill, Heidur Gretarsdottir, Sigridur Sigurdardottir, Kristin Anna Olafsdottir, Anna Sigridur Islind, María Óskarsdóttir, et al. Multicentre sleep-stage scoring agreement in the sleep revolution project. *Journal of Sleep Research*, 33 (1):e13956, 2024.
21. Henna Pitkänen, Sami Nikkonen, Marika Rissanen, Anna Sigridur Islind, Heidur Gretarsdottir, Erna Sif Arnardottir, Timo Leppänen, and Henri Korkalainen. Multi-centre arousal scoring agreement in the sleep revolution. *Journal of Sleep Research*, 33(4):e14127, 2024.
22. Minna Pitkänen, Henna Pitkänen, Rajdeep Kumar Nath, Sami Nikkonen, Samu Kainulainen, Henri Korkalainen, Kristín Anna Ólafsdóttir, Erna Sif Arnardottir, Sigridur Sigurdardottir, Thomas Penzel, et al. Temporal and sleep stage-dependent agreement in manual scoring of respiratory events. *Journal of Sleep Research*, page e14391, 2024.
23. Rechtschaffen A and Kales A. A manual of standardized termi- nology, techniques and scoring system of sleep stages in human subjects. *U.S. Public Health Service*, 1968.
24. RB Berry, R Brooks, CE Gamaldo, SM Harding, RM Lloyd, CL Marcus, and BV Vaughn. The aasm manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. 3, 2023.
25. Sari-Leena Himanen and Joel Hasan. Limitations of rechtschaffen and kales. *Sleep medicine reviews*, 4(2): 149–167, 2000.
26. Mohammadreza Ghorvei, Tuomas Karhu, Salla Hietakoste, Daniela Ferreira-Santos, Harald Hrubos-Strøm, Anna Sigridur Islind, Luka Biedebach, Sami Nikkonen, Timo Leppänen, and Matias Rusanen. A comparative analysis of unsupervised machine-learning methods in psg-related phenotyping. *Journal of Sleep Research*, page e14349, 2024.
27. Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.
28. Lior Rokach and Oded Maimon. Clustering methods. *Data mining and knowledge discovery handbook*, pages 321–352, 2005.
29. John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.
30. James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203, 1984.

31. Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.
32. Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(3):231–240, 2011.
33. Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.
34. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
35. Ian Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.
36. Aapo Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110534, 2013.
37. Sotiris Kotsiantis and Dimitris Kanellopoulos. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32 (1):71–82, 2006.
38. Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A Inkeri Verkamo, et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1):307–328, 1996.
39. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
40. Lawrence Rabiner and Biinghwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.
41. Sean R Eddy. What is a hidden markov model? *Nature biotechnology*, 22(10):1315–1316, 2004.
42. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection for discrete sequences: A survey. *IEEE transactions on knowledge and data engineering*, 24(5): 823–839, 2010.
43. Shikha Agrawal and Jitendra Agrawal. Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60:708–713, 2015.
44. Luka Biedebach, María Óskarsdóttir, Erna Sif Arnardóttir, Sigridur Sigurdardóttir, Michael Valur Clausen, Sigurveig Þ Sigurdardóttir, Marta Serwatko, and Anna Sigridur Islind. Anomaly detection in sleep: detecting mouth breathing in children. *Data Mining and Knowledge Discovery*, 38(3):976–1005, 2024.
45. Mai T Pham, Andrijana Rajić, Judy D Greig, Jan M Sargeant, Andrew Papadopoulos, and Scott A McEwen. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Research synthesis methods*, 5(4):371–385, 2014.
46. Andrea C Tricco, Erin Lillie, Wasifa Zarin, Kelly K O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah DJ Peters, Tanya Horsley, Laura Weeks, et al. Prisma extension for scoping reviews (prisma-scr): checklist and explanation. *Annals of internal medicine*, 169(7):467–473, 2018.
47. Luka Biedebach. Unsupervised machine learning in sleep research - a scoping review protocol, Jul 2024. URL `osf.io/42zrb`.
48. Gabriel Natan Pires, Erna S Arnardóttir, Jose M Saavedra, Sergio Tufik, and Walter T McNicholas. Search filters for systematic reviews and meta-analyses in sleep medicine. *Sleep Medicine*, 2025.
49. Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5:1–10, 2016.
50. Absalom E Ezugwu, Japie Greeff, and Yuh-Shan Ho. A comprehensive study of groundbreaking machine learning research: analyzing highly cited and impactful publications across six decades. *Journal of Engineering Research*, 2023.
51. A. Kumar. A real-time system for pattern recognition of human sleep stages by fuzzy system analysis. *Pattern Recognition*, 9(1):43–46, 1977.
52. I. Gath and E. Bar-On. Classical sleep stages and the spectral content of the EEG signal. *The International journal of neuroscience*, 22(1):147–155, December 1983.
53. I. Gath and L. Schwartz. Syntactic pattern recognition applied to sleep EEG staging. *Pattern Recognition Letters*, 10(4):265–272, 1989.
54. I. Gath and E. Bar-on. Computerized method for scoring of polygraphic sleep recordings. *Computer programs in biomedicine*, 11(3):217–223, June 1980.
55. I. Gath and A.B. Geva. Unsupervised Optimal Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–780, 1989.
56. I. Gath, C. Feuerstein, and A. Geva. Unsupervised classification and adaptive definition of sleep patterns. *Pattern Recognition Letters*, 15(10):977–984, 1994.
57. B. H. Jansen and W. K. Cheng. Classification of sleep patterns by means of markov modeling and correspondence analysis. *IEEE transactions on pattern analysis and machine intelligence*, 9(5):707–710, May 1987.
58. H. Escola, E. Poiseau, M. Jobert, and P. Gaillard. Classification using distance-based segmentation-application to the analysis of EEG signals. *Pattern Recognition Letters*, 12(6):327–333, 1991.
59. T. Katayama, E. Suzuki, and M. Saito. Staging of awake and sleep based on feature map. *Systems and Computers in Japan*, 26(7):98–107, 1995.
60. H. Larsen and D.C. Lai. Walsh Spectral Estimates with Applications to the Classification of EEG Signals. *IEEE Transactions on Biomedical Engineering*, (9):485–492, 1980.
61. J. Kohlmorgen, K.-R. Müllerc, and K. Pawelzik. Analysis of drifting dynamics with neural network hidden markov models. *Advances in Neural Information Processing Systems*, pages 735–741, 1998.
62. R. Rosipal, G. Dorffner, and E. Trenker. Can ICA improve sleep-spindles detection? *Neural Network World*, 8(5): 539–547, 1998.
63. O. Caspary, M. Tomczak, N. Di Renzo, M. Mouze-Amady, and D. Henry. Enhanced high resolution spectral analysis of sleep spindles. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 16:1232–1233, 1994.
64. P. Lima, J. Leitao, and T. Paiva. Artifact detection in sleep EEG recording. pages 273–277, 1989.
65. Osvaldo Rocha Pacheco and Francisco Vaz. Integrated system for analysis and automatic classification of sleep EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 4: 2062–2065, 1998.

66. V.L. Ramnath and S. Katkoori. A Smart IoT System for Continuous Sleep State Monitoring. *Midwest Symposium on Circuits and Systems*, 2020:241–244, 2020.
67. Q. Pan, D. Brulin, E. Campo, and Patnaik S. Home sleep monitoring based on wrist movement data processing. *Procedia Computer Science*, 183:696–705, 2021.
68. Arun Sebastian, Peter Cistulli, Gary Cohen, and Philip de Chazal. A Preliminary Study of the Automatic Classification of the Site of Airway Collapse in OSA patients Using Snoring Signals. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2019: 1592–1595, July 2019.
69. I. Takao, K. Nishio, T. Kaburagi, S. Kumagai, T. Matsumoto, and Y. Kurihara. A Home Sleep Apnea State Monitoring System using a Stacked Autoencoder. *Proceedings of IEEE Sensors*, 2019, 2019.
70. A.L. Alfeo, P. Barsocchi, M.G.C.A. Cimino, D. La Rosa, F. Palumbo, and G. Vaglini. Sleep behavior assessment via smartwatch and stigmergic receptive fields. *Personal and Ubiquitous Computing*, 22(2):227–243, 2018.
71. Z. Liang, B. Ploderer, M.A.C. Martell, T. Nishimura, Martin-Gonzalez A, and Uc-Cetina V. A cloud-based intelligent computing system for contextual exploration on personal sleep-tracking data using association rule mining. *Communications in Computer and Information Science*, 597:83–96, 2016.
72. Z. Liang, M.A.C. Martell, and T. Nishimura. Mining hidden correlations between sleep and lifestyle factors from quantified-self data. *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 547–552, 2016.
73. Luka Biedebach, María Óskarsdóttir, Erna Sif Arnardottir, and Anna Sigríður Islind. Two Sides of the Same Pillow: Unfolding the Relationship between Objective and Subjective Sleep Quality with Unsupervised Learning. 2023.
74. I.W. Muns, Y. Lad, I.G. Guardiola, M. Thimgan, and Dagli C.H. Classification of Rest and Active Periods in Actigraphy Data Using PCA. *Procedia Computer Science*, 114:275–280, 2017.
75. Y. El-Manzalawy, O. Buxton, V. Honavar, Yoo I, Zheng J.H, Gong Y, Hu X.T, Shyu C.-R, Bromberg Y, Gao J, and Korkin D. Sleep/wake state prediction and sleep parameter estimation using unsupervised classification via clustering. *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, 2017:718–723, 2017.
76. Yan Ren, Siva Sivaganesan, Mekibib Altaye, Raouf S. Amin, and Rhonda D. Szczesniak. Biclustering of medical monitoring data using a nonparametric hierarchical Bayesian model. *Stat (International Statistical Institute)*, 9(1):e279, 2020.
77. A. Gasmi, V. Augusto, J. Faucheu, C. Morin, and X. Serpaggi. Anomaly Detection in Sleep Habits Using Deep Learning. *IEEE International Conference on Automation Science and Engineering*, 2023, 2023.
78. Stijn A. A. Massar, Xin Yu Chua, Chun Siong Soon, Alyssa S. C. Ng, Ju Lynn Ong, Nicholas I. Y. N. Chee, Tih Shih Lee, Arko Ghosh, and Michael W. L. Chee. Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data. *NPJ digital medicine*, 4 (1):90, June 2021.
79. Yuning Wang, Iman Azimi, Mohammad Feli, Amir M. Rahmani, and Pasi Liljeberg. Personalized Graph Attention Network for Multivariate Time-series Change Analysis: A Case Study on Long-term Maternal Monitoring. pages 593–598, 2023.
80. L. Zhang, Q.F. Zhou, and M. Peng. Intelligent mattress aliasing signal decomposition based on SVD. *ACM International Conference Proceeding Series*, pages 201–205, 2019.
81. S. Ostadabbas, M. Baran Pouyan, M. Nourani, and N. Kehtarnavaz. In-bed posture classification and limb identification. *IEEE 2014 Biomedical Circuits and Systems Conference, BioCAS 2014 - Proceedings*, pages 133–136, 2014.
82. M. Baran Pouyan, M. Nourani, and M. Pompeo. Clustering-based limb identification for pressure ulcer risk assessment. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2015:4230–4233, 2015.
83. A.M. Adami, M. Pavel, T.L. Hayes, A.G. Adami, and C. Singer. A method for classification of movements in bed. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 7881–7884, 2011.
84. M. Uchida, W. Chen, T. Nemoto, K. Kitamura, Y. Kanemitsu, and D. Wei. An ICA approach to reject noise from pressure changes of pillow. *Proceedings - The Fourth International Conference on Computer and Information Technology (CIT 2004)*, pages 916–921, 2004.
85. Rong-Shue Hsiao, Tian-Xiang Chen, Mekuanint Agegnehu Bitew, Chun-Hao Kao, and Tzu-Yu Li. Sleeping posture recognition using fuzzy c-means algorithm. *Biomedical engineering online*, 17:157, November 2018.
86. S. Bhatlawande and S. Kulkarni. Residential Monitoring System for Classification and Recognition of Sleeping Posture. *2022 2nd International Conference on Intelligent Technologies, CONIT 2022*, 2022.
87. Y. Hou and F. Zhu. Contactless heart rate and respiratory measurement for sleep monitoring. *Proceedings of the 12th EAI International Conference on Mobile Multimedia Communications, MOBIMEDIA 2019*, pages 199–211, 2019.
88. Kaiyin Zhu, Michael Li, Sina Akbarian, Maziar Hafezi, Azadeh Yadollahi, and Babak Taati. Vision-Based Heart and Respiratory Rate Monitoring During Sleep - A Validation Study for the Population at Risk of Sleep Apnea. *IEEE journal of translational engineering in health and medicine*, 7:1900708, 2019.
89. N. Koolen, O. Decroupet, A. Dereymaeker, K. Jansen, J. Vervisch, V. Matic, B. Vanrumste, G. Naulaers, S. Van Huffel, M. De Vos, De Marsico M, Figueiredo M, and Fred A. Automated respiration detection from neonatal video data. *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings*, 2:164–169, 2015.
90. Kh. Rezaee, Gh. Mohammadi, I. Mirzajani, and J. Hadadnia. Real-time intelligent alarm system of driver fatigue based on video sequences. *Iran Occupational Health*, 10(3):1–11, 2013.
91. C. Schwarz, J. Gaspar, and R. Yousefian. Sequence Analysis of Monitored Drowsy Driving. *Transportation*

Research Record, 2677(8):553–562, 2023.

92. Jianan Han, Shaoxing Zhang, Aidong Men, and Qingchao Chen. Cross-Modal Contrastive Hashing Retrieval for Infrared Video and EEG. *Sensors (Basel, Switzerland)*, 22(22), November 2022.
93. D. Falie and M. Ichim. Statistical signal analysis in Lp space. *Proceedings - 2010 3rd International Congress on Image and Signal Processing, CISP 2010*, 7:3173–3177, 2010.
94. Filipe Barata, Peter Tinschert, Frank Rassouli, Claudia Steurer-Stey, Elgar Fleisch, Milo Alan Puhan, Martin Brutsche, David Kotz, and Tobias Kowatsch. Automatic Recognition, Segmentation, and Sex Assignment of Nocturnal Asthmatic Coughs and Cough Epochs in Smartphone Audio Recordings: Observational Field Study. *Journal of medical Internet research*, 22(7):e18082, July 2020.
95. Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller. Snore-GANs: Improving Automatic Snore Sound Classification with Synthesized Data. *IEEE Journal of Biomedical and Health Informatics*, 24(1): 300–310, 2020.
96. R. J. Beeton, I. Wells, P. Ebden, H. B. Whittet, and J. Clarke. Snore site discrimination using statistical moments of free field snoring sounds recorded during sleep nasendoscopy. *Physiological measurement*, 28(10): 1225–1236, October 2007.
97. N. Ben-Israel, A. Tarasiuk, and Y. Zigel. Nocturnal sound analysis for the diagnosis of obstructive sleep apnea. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, pages 6146–6149, 2010.
98. Y. Zhao, H. Zhang, W. Liu, and S. Ding. A snoring detector for OSAHS based on patient's individual personality. *Proceedings of 2011 3rd International Conference on Awareness Science and Technology, iCAST 2011*, pages 24–27, 2011.
99. Evgenia Goldshtein, Ariel Tarasiuk, and Yaniv Zigel. Automatic detection of obstructive sleep apnea using speech signals. *IEEE transactions on bio-medical engineering*, 58(5):1373–1382, May 2011.
100. K.-I. Fukui, S. Ishimaru, T. Kato, and M. Numao. Sound-based sleep assessment with controllable subject-dependent embedding using Variational Domain Adversarial Neural Network. *International Journal of Data Science and Analytics*, 2023.
101. J.L. Blanco, L.A. Hernández, R. Fernández, and D. Ramos. Improving Automatic Detection of Obstructive Sleep Apnea Through Nonlinear Analysis of Sustained Speech. *Cognitive Computation*, 5(4):458–472, 2013.
102. J.-A. Gómez-García, J.-L. Blanco-Murillo, J.-I. Godino-Llorente, L.A. Hernández Gómez, and G. Castellanos-Domínguez. GMM-based classifiers for the automatic detection of Obstructive Sleep Apnea. *BIOSIGNALS 2013 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, pages 364–367, 2013.
103. R. Fernández, J.L. Blanco, L.A. Hernández, E. López, J. Alcázar, and D.T. Toledano. Severe APNOEA detection using speaker recognition techniques. *BIOSIGNALS 2009 - Proceedings of the 2nd International Conference on Bio-Inspired Systems and Signal Processing*, pages 124–130, 2009.
104. Z.K. Shahid, S. Saguna, and C. Ahlund. Recognizing Long-term Sleep Behaviour Change using Clustering for Elderly in Smart Homes. *ISC2 2022 - 8th IEEE International Smart Cities Conference*, 2022.
105. R. Obukata, M. Cuka, D. Elmazi, T. Oda, K. Matsuo, and L. Barolli. Implementation of an actor node for an ambient intelligence testbed considering bed temperature and room lighting: Its effects on human sleeping condition. *Lecture Notes on Data Engineering and Communications Technologies*, 8:73–81, 2018.
106. Y Gu, YF Zhang, J Li, YS Ji, X An, and FJ Ren. Sleepy: Wireless Channel Data Driven Sleep Monitoring via Commodity WiFi Devices. *IEEE TRANSACTIONS ON BIG DATA*, 6(2):258–268, June 2020.
107. Yee Siong Lee, Pubudu N. Pathirana, Robin J. Evans, and Christopher L. Steinfort. Separation of Doppler radar-based respiratory signatures. *Medical & biological engineering & computing*, 54(8):1169–1179, August 2016.
108. A. Caroppo, A. Leone, G. Rescio, G. Diraco, P. Siciliano, Siciliano P, Di Natale C, Baldini F, Ando B, and Marrazza G. Multi-sensor platform for detection of anomalies in human sleep patterns. *Lecture Notes in Electrical Engineering*, 431:276–285, 2018.
109. C.F. Goh, L.B. Samuclsson, M.H. Hall, G.G. Lee Seet, and K. Shimada. Semi-Automatic Snore Detection in Polysomnography based on Hierarchical Clustering. *IEEE International Conference on Automation Science and Engineering*, 2018:1116–1122, 2018.
110. M.F. Bagci, T. Nguyen, and Y. Ozturk. Ambient Sleep Quality Analysis with a Machine Learning Model. *ICASSPW 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing Workshops, Proceedings*, 2023.
111. J.-J. Chung and H.-J. Kim. An automobile environment detection system based on deep neural network and its implementation using iot-enabled in-vehicle air quality sensors. *Sustainability (Switzerland)*, 12(6), 2020.
112. M.K. Wali, M. Murugappan, and R. Badlishah Ahmad. Classification of driver drowsiness level using wireless EEG. *Przeglad Elektrotechniczny*, 89(6):113–117, 2013.
113. Kyle Q. Lepage, Sparsh Jain, Andrew Kvavilashvili, Mark Witcher, and Sujith Vijayan. Unsupervised Multitaper Spectral Method for Identifying REM Sleep in Intracranial EEG Recordings Lacking EOG/EMG Data. *Bioengineering (Basel, Switzerland)*, 10(9), August 2023.
114. Samantha Sun, Linxing Preston Jiang, Steven M. Peterson, Jeffrey Herron, Kurt Weaver, Andrew Ko, Jeffrey Ojemann, and Rajesh P. N. Rao. Unsupervised Sleep and Wake State Identification in Long-Term Electrocorticography Recordings. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2020: 629–632, July 2020.
115. Jong-Hwan Lee, Sungsuk Oh, Ferenc A. Jolesz, Hyunwook Park, and Seung-Schik Yoo. Application of independent component analysis for the data mining of simultaneous Eeg-fMRI: preliminary experience on sleep onset. *The International journal of neuroscience*, 119(8):1118–1136, 2009.
116. Evan Houldin, Zhuo Fang, Laura B. Ray, Adrian M. Owen, and Stuart M. Fogel. Toward a complete taxonomy of resting state networks across wakefulness and sleep: an assessment of spatially distinct resting state networks

using independent component analysis. *Sleep*, 42(3): zsy235, March 2019.

117. Trung Q. Le, Changqing Cheng, Akkarapol Sangasoongsong, Woranat Wongdhamma, and Satish T. S. Bukkapatnam. Wireless Wearable Multisensory Suite and Real-Time Prediction of Obstructive Sleep Apnea Episodes. *IEEE journal of translational engineering in health and medicine*, 1:2700109, 2013.
118. C. Fung, J. Lopez, S. Hurtado, E. Garcia, and Velazquez R. Intelligent Lighting System Based on Sleep Detection Using IoT Devices and Wearables. *Proceedings - 2023 IEEE Latin-American Conference on Communications, LATINCOM 2023*, 2023.
119. Jiaxing Liu, Yang Zhao, Boya Lai, Hailiang Wang, and Kwok Leung Tsui. Wearable Device Heart Rate and Activity Data in an Unsupervised Approach to Personalized Sleep Monitoring: Algorithm Validation. *JMIR mHealth and uHealth*, 8(8):e18370, August 2020.
120. D. Geng, Z. Qin, J. Wang, Z. Gao, and N. Zhao. Personalized recognition of wake/sleep state based on the combined shapelets and K-means algorithm. *Biomedical Signal Processing and Control*, 71, 2022.
121. K. Fujiwara, H. Iwamoto, K. Hori, and M. Kano. Driver Drowsiness Detection Using R-R Interval of Electrocardiogram and Self-Attention Autoencoder. *IEEE Transactions on Intelligent Vehicles*, pages 1–10, 2023.
122. I. Hermawan, M.I. Tawakal, I.M.A. Setiawan, I. Habibie, and W. Jatmiko. Adaptive Multi codebook Fuzzy Neuro Generalized Learning Vector Quantization for sleep stages classification. *2013 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2013*, pages 431–436, 2013.
123. M. Takahashi, N. Sugahara, and M. Shibata. Towards detecting morning surge from sleep self-evaluations. *Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2020*, pages 1–6, 2020.
124. Villads Hulgaard Joergensen, Umaer Hanif, Poul Jennum, Emmanuel Mignot, Asbjoern W. Helge, and Helge B. D. Sorensen. Automatic Segmentation to Cluster Patterns of Breathing in Sleep Apnea. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2021: 164–168, November 2021.
125. H. Elmoaqet, J. Kim, D. Tilbury, S.K. Ramachandran, M. Ryalat, and C.-H. Chu. Gaussian mixture models for detecting sleep apnea events using single oronasal airflow record. *Applied Sciences (Switzerland)*, 10(21):1–15, 2020.
126. Tero Aittokallio, Jani S. Malminen, Tapio Pahikkala, Olli Polo, and Olli S. Nevalainen. Inspiratory flow shape clustering: an automated method to monitor upper airway performance during sleep. *Computer methods and programs in biomedicine*, 85(1):8–18, January 2007.
127. P. Loliencar and G. Heo. Phenotyping OSA: a time series analysis using fuzzy clustering and persistent homology. *International Journal of Approximate Reasoning*, 142: 178–195, 2022.
128. R. Haidar, I. Koprinska, B. Jeffries, Gedeon T, Wong K.W, and Lee M. Feature learning and data compression of biosignals using convolutional autoencoders for sleep apnea detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11953:162–174, 2019.
129. B. Holm, A.S. Islind, E.S. Arnardóttir, M. Óskarsdóttir, and Bui T.X. Exploration of Sleep Events in the Latent Space of Variational Autoencoders on a Breath-by-Breath Basis. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2023:3091–3100, 2023.
130. J. Vanbuis, M. Feuilloy, G. Baffet, N. Meslier, F. Gagnadoux, and J.-M. Girault. A New Sleep Staging System for Type III Sleep Studies Equipped with a Tracheal Sound Sensor. *IEEE Transactions on Biomedical Engineering*, 69(3):1225–1236, 2022.
131. U. Reimer, S. Emmenegger, E. Maier, T. Ulmer, H.-J. Vollbrecht, Z. Zhang, R. Khatami, Maciaszek L, O. Donoghue J, Molloy W, Rocker C, and Ziefle M. Laying the foundation for correlating daytime behaviour with sleep architecture using wearable sensors. *Communications in Computer and Information Science*, 869:147–167, 2018.
132. S.S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García. SpO2 based sleep apnea detection using deep learning. *INES 2017 - IEEE 21st International Conference on Intelligent Engineering Systems, Proceedings*, 2017:91–96, 2017.
133. J. Víctor Marcos, Roberto Hornero, Daniel Alvarez, Félix del Campo, Miguel López, and Carlos Zamarrón. Radial basis function classifiers to help in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry. *Medical & biological engineering & computing*, 46(4):323–332, April 2008.
134. Z. Li, M. Arvaneh, H.E. Elphick, R.N. Kingshott, and L.S. Mihaylova. A dirichlet process mixture model for autonomous sleep apnea detection using oxygen saturation data. *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020*, 2020.
135. Malak Abdullah Almarshad, Saad Al-Ahmadi, Md Saiful Islam, Ahmed S. BaHammam, and Adel Soudani. Adoption of Transformer Neural Network to Improve the Diagnostic Performance of Oximetry for Obstructive Sleep Apnea. *Sensors (Basel, Switzerland)*, 23(18), September 2023.
136. M.S. Daley, K. Diaz, H.F. Posada-Quintero, Y. Kong, K. Chon, and J.B. Bolkhovsky. Archetypal physiological responses to prolonged wakefulness. *Biomedical Signal Processing and Control*, 74, 2022.
137. S. Usami. Constrained k-means on cluster proportion and distances among clusters for longitudinal data analysis. *Japanese Psychological Research*, 56(4):361–372, 2014.
138. E.-Y. Ma, J.-W. Kim, Y. Lee, S.-W. Cho, H. Kim, and J.K. Kim. Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea. *Scientific Reports*, 11(1), 2021.
139. Hyeonhoon Lee, Yujin Choi, Byunwoo Son, Jinwoong Lim, Seunghoon Lee, Jung Won Kang, Kun Hyung Kim, Eun Jung Kim, Changsop Yang, and Jae-Dong Lee. Deep autoencoder-powered pattern identification of sleep disturbance using multi-site cross-sectional survey data. *Frontiers in medicine*, 9:950327, 2022.
140. M.L.U.R. Shahid, V. Molchanov, J. Mir, F. Shaukat, and L. Linsen. Interactive visual analytics tool for multidimensional quantitative and categorical data analysis. *Information Visualization*, 19(3):234–246, 2020.
141. J. Hong and J. Yoon. Multivariate time-series classification of sleep patterns using a hybrid deep learning architecture.

2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom 2017, 2017:1–6, 2017.

142. K.A. Wong, A. Paul, P. Fuentes, D.C. Lim, A. Das, and M. Tan. Screening for obstructive sleep apnea in patients with cancer - A machine learning approach. *SLEEP Advances*, 4(1), 2023.
143. S.T. Mai, S. Amer-Yahia, S. Bailly, J.-L. Pépin, A.D. Chouakria, K.T. Nguyen, and A.-D. Nguyen. Evolutionary Active Constrained Clustering for Obstructive Sleep Apnea Analysis. *Data Science and Engineering*, 3(4): 359–378, 2018.
144. Joshua R. Mirth, Christopher L. Felton, Clifton R. Haider, Stuart J. McCarter, Timothy I. Morgenthaler, Erik K. St Louis, and David R. Holmes. Identification of Sleep Patterns via Clustering of Hypnodensities. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2023:1–4, July 2023.
145. 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.
146. M.K.R. Baddam, M. Araujo, J. Srivastava, Almeida J.R, Gonzalez A.R, Shen L, Kane B, Traina A, Soda P, and Oliveira J.L. Defining and monitoring patient clusters based on therapy adherence in sleep apnea management. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2021:580–585, 2021.
147. Y. Kang, V.V. Prabhu, A.M. Sawyer, and P.M. Griffin. Markov models for treatment adherence in obstructive sleep apnea. *IIE Annual Conference and Expo 2013*, pages 1592–1599, 2013.
148. P. Boyraz, M. Acar, and D. Kerr. Multi-sensor driver drowsiness monitoring. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 222(11):1857–1878, 2008.
149. Zuzana Rošťáková and Roman Rosipal. Profiling continuous sleep representations for better understanding of the dynamic character of normal sleep. *Artificial intelligence in medicine*, 97:152–167, June 2019.
150. L.M. Sepulveda-Cano, E. Gil, P. Laguna, and G. Castellanos-Dominguez. Selection of nonstationary dynamic features for obstructive sleep apnoea detection in children. *Eurasip Journal on Advances in Signal Processing*, 2011, 2011.
151. Ran Wei, Xinghua Zhang, Jinhai Wang, and Xin Dang. The research of sleep staging based on single-lead electrocardiogram and deep neural network. *Biomedical engineering letters*, 8(1):87–93, February 2018.
152. S. Kim, D. Lee, H. Kwak, and S. Lee. Towards Domain-free Transformer for Generalized EEG Pre-training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1–1, 2024.
153. Sudhanshu S. D. P. Ayyagari, Richard D. Jones, and Stephen J. Weddell. Detection of microsleep states from the EEG: a comparison of feature reduction methods. *Medical & biological engineering & computing*, 59(7): 1643–1657, August 2021.
154. S.A. Alvarez and C. Ruiz. Collective probabilistic dynamical modeling of sleep stage transitions. *BIOSIGNALS 2013 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, pages 209–214, 2013.
155. F. Mendonça, S.S. Mostafa, F. Morgado-Dias, and A.G. Ravelo-García. Cyclic alternating pattern estimation based on a probabilistic model over an EEG signal. *Biomedical Signal Processing and Control*, 62, 2020.
156. Y. Zhang, J. Wang, Y. Chen, H. Yu, and T. Qin. Adaptive Memory Networks With Self-Supervised Learning for Unsupervised Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12068–12080, 2023.
157. Koichi Fujiwara, Erika Abe, Keisuke Kamata, Chikao Nakayama, Yoko Suzuki, Toshitaka Yamakawa, Toshihiro Hiraoka, Manabu Kano, Yukiyoshi Sumi, Fumi Masuda, Masahiro Matsuo, and Hiroshi Kadotani. Heart Rate Variability-Based Driver Drowsiness Detection and Its Validation With EEG. *IEEE transactions on bio-medical engineering*, 66(6):1769–1778, June 2019.
158. H. Wang, A. Li, T. Chen, J. Liu, and Shi Y. Study on behavioral risk for aging based on smart mattress monitoring data. *Procedia Computer Science*, 221: 1276–1283, 2023.
159. Miguel R. Álvarez, Paulo Félix, and Purificación Cariñena. Discovering metric temporal constraint networks on temporal databases. *Artificial intelligence in medicine*, 58(3):139–154, July 2013.
160. P. Laxminarayan, S.A. Alvarez, C. Ruiz, and M. Moonis. Mining associations over human sleep time series. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, pages 323–325, 2005.
161. J.-C. Kim and K. Chung. Mining Based Time-Series Sleeping Pattern Analysis for Life Big-Data. *Wireless Personal Communications*, 105(2):475–489, 2019.
162. Parameshvyas Laxminarayan, Sergio A. Alvarez, Carolina Ruiz, and Majaz Moonis. Mining statistically significant associations for exploratory analysis of human sleep data. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 10(3):440–450, July 2006.
163. Rashmie Abeysinghe and Licong Cui. Query-constraint-based mining of association rules for exploratory analysis of clinical datasets in the National Sleep Research Resource. *BMC medical informatics and decision making*, 18:58, July 2018.
164. C.A. Loza, L.L. Colgin, de Bruijne M, Cattin P.C, Cotin S, Padoy N, Speidel S, Zheng Y, and Essert C. Deep Neural Dynamic Bayesian Networks Applied to EEG Sleep Spindles Modeling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12905:550–560, 2021.
165. Carlos Miguel Fernandes, Antonio Mora, Juan Julian Merelo, Francisco Fernández, and Agostinho Rosa. Generating colored 2-dimensional representations of sleep EEG with the KANTS clustering algorithm. pages 435–442, 2012.
166. S. Kumi, M. Hilton, C. Snow, R.K. Lomotey, R. Deters, Chang C.K, Chang R.N, Fan J, Fox G.C, Jin Z, Pravadelli G, and Shahriar H. SleepSynth: Evaluating the use of Synthetic Data in Health Digital Twins. *Proceedings - 2023 IEEE International Conference on Digital Health, ICDH 2023*, pages 121–130, 2023.
167. C.A. Loza, J.C. Principe, Joya G, Rojas I, and Catala A. The Generalized Sleep Spindles Detector: A Generative

Model Approach on Single-Channel EEGs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11506:127–138, 2019.

168. Toki Takeda, Osamu Mizuno, and Tomohiro Tanaka. Time-dependent sleep stage transition model based on heart rate variability. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2015:2343–2346, 2015.
169. L. Biedebach, M. Rusanen, B.H. Pórðarson, E.S. Arnardóttir, M. Óskarsdóttir, S. Nikkonen, H. Korkalainen, S. Myllymaa, J. Töyräs, S. Kainulainen, T. Leppänen, and A.S. Islind. Towards a Deeper Understanding of Sleep Stages through their Representation in the Latent Space of Variational Autoencoders. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2023: 3111–3120, 2023.
170. E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, X. Li, and C. Guan. ADAST: Attentive Cross-Domain EEG-Based Sleep Staging Framework With Iterative Self-Training. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(1):210–221, 2023.
171. Dong-Rui Gao, Jing Li, Man-Qing Wang, Lu-Tao Wang, and Yong-Qing Zhang. Automatic sleep staging of single-channel EEG based on domain adversarial neural networks and domain self-attention. *Frontiers in neuroscience*, 17: 1143495, 2023.
172. Z. He, M. Tang, P. Wang, L. Du, X. Chen, G. Cheng, and Z. Fang. Cross-scenario automatic sleep stage classification using transfer learning and single-channel EEG. *Biomedical Signal Processing and Control*, 81, 2023.
173. Elisabeth R. M. Heremans, Huy Phan, Pascal Borzée, Bertien Buyse, Dries Testelmans, and Maarten De Vos. From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging. *Journal of neural engineering*, 19(3), June 2022.
174. W. Qu, C.-H. Kao, H. Hong, Z. Chi, R. Grunstein, C. Gordon, and Z. Wang. Single-channel EEG based insomnia detection with domain adaptation. *Computers in Biology and Medicine*, 139, 2021.
175. C. Yoo, H.W. Lee, and J.-W. Kang. Transferring Structured Knowledge in Unsupervised Domain Adaptation of a Sleep Staging Network. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1273–1284, 2022.
176. Y. Luo, Y. Zheng, H. Shao, L. Zhang, and L. Li. TUDAMatch: Time-Series Unsupervised Domain Adaptation for Automatic Sleep Staging. *International IEEE/EMBS Conference on Neural Engineering, NER*, 2023, 2023.
177. J. Fan, H. Zhu, X. Jiang, L. Meng, C. Chen, C. Fu, H. Yu, C. Dai, and W. Chen. Unsupervised Domain Adaptation by Statistics Alignment for Deep Sleep Staging Networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:205–216, 2022.
178. R. Zhao, Y. Xia, and Y. Zhang. Unsupervised sleep staging system based on domain adaptation. *Biomedical Signal Processing and Control*, 69, 2021.
179. J. Zhang and Y. Wu. Competition convolutional neural network for sleep stage classification. *Biomedical Signal Processing and Control*, 64, 2021.
180. Y. Amor, L. Rejeb, R. Ferjeni, L. Ben Said, and M.R. Ben Cheikh. Hierarchical Multi-agent System for Sleep Stages Classification. *International Journal on Artificial Intelligence Tools*, 31(5), 2022.
181. Y. El-Khadiri, G. Corona, C. Rose, and F. Charpillet. Sleep activity recognition using binary motion sensors. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2018:265–269, 2018.
182. Q. Xiao, J. Wang, J. Ye, H. Zhang, Y. Bu, Y. Zhang, and H. Wu. Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2021:1290–1294, 2021.
183. S. Güneş, K. Polat, and S. Yosunkaya. Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Systems with Applications*, 37(12):7922–7928, 2010.
184. K. Yin, R. Zhu, S. Hou, and G. Yin. Unsupervised Assisted Sleep staging Classification Algorithm under Fuzzy Few Samples. *11th International Conference on Intelligent Control and Information Processing, ICICIP 2021*, pages 154–159, 2021.
185. I. Mporas, A. Efstathiou, V. Megalooikonomou, Guo Y. Y, Hill S. S, Friston K, Peng H, and Faisal A. A. Sleep stages classification from electroencephalographic signals based on unsupervised feature space clustering. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9250:77–85, 2015.
186. P. Tian, J. Hu, J. Qi, X. Ye, D. Che, Y. Ding, and Y. Peng. A hierarchical classification method for automatic sleep scoring using multiscale entropy features and proportion information of sleep architecture. *Biocybernetics and Biomedical Engineering*, 37(2):263–271, 2017.
187. V. Krajca, H. Schaabova, V. Piorecka, M. Piorecky, J. Strobl, L. Lhotska, V. Gerla, K. Paul, Lhotska L, Sukupova L, Lackovic I, and Ibbott G.S. Detection of sleep stages in temporal profiles in neonatal EEG—k-NN versus k-means approach: A feasibility study. *IFMBE Proceedings*, 68(2):523–527, 2018.
188. Yi-Feng Ko, Pei-Hsin Kuo, Ching-Fu Wang, Yu-Jen Chen, Pei-Chi Chuang, Shih-Zhang Li, Bo-Wei Chen, Fu-Chi Yang, Yu-Chun Lo, Yi Yang, Shuan-Chu Vina Ro, Fu-Shan Jaw, Sheng-Huang Lin, and You-Yin Chen. Quantification Analysis of Sleep Based on Smartwatch Sensors for Parkinson's Disease. *Biosensors*, 12(2), January 2022.
189. M.L. Trevenen, B.A. Turlach, P.R. Eastwood, L.M. Straker, and K. Murray. Using hidden Markov models with raw, triaxial wrist accelerometry data to determine sleep stages. *Australian and New Zealand Journal of Statistics*, 61(3):273–298, 2019.
190. Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.
191. Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.
192. K. Rai, V. Bajaj, and A. Kumar. Hilbert-Huang transform based classification of sleep and wake EEG signals using

fuzzy c-means algorithm. *2015 International Conference on Communication and Signal Processing, ICCSP 2015*, pages 460–464, 2015.

193. Z.K. Shahid, S. Saguna, and C. Ahlund. Multi-Armed Bandits for Sleep Recognition of Elderly Living in Single-Resident Smart Homes. *IEEE Internet of Things Journal*, pages 1–1, 2023.
194. Sandya Subramanian and Todd P. Coleman. Automated classification of sleep and wake from single day triaxial accelerometer data. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2022:3665–3668, July 2022.
195. Iris A. M. Huijben, Arthur A. Nijdam, Lieke W. A. Hermans, Sebastiaan Overeem, Merel M. Van Gilst, and Ruud J. G. Van Sloun. Self-Organizing Maps for Contrastive Embeddings of Sleep Recordings. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2022:2945–2948, July 2022.
196. J. Li, X. Ge, L. Hu, Q. Zhu, Z. Zhang, and F. Lv. Measuring Sleep Stages and Quality Based on K-Means and Random Forest Algorithms. *2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2023*, pages 484–488, 2023.
197. R. Alabdan, H.A. Mengash, M. Maray, F. Alotaibi, S. Abdelbagi, and A. Mahmud. Modified Bald Eagle Search Algorithm with Deep Learning-Driven Sleep Quality Prediction for Healthcare Monitoring Systems. *IEEE Access*, 11:135385–135393, 2023.
198. P. Khumngoen and S. Sinthupinyo. Sleep Behavior Classification Based on Clusters of Sleep Quality. *Proceedings of JCSSE 2023 - 20th International Joint Conference on Computer Science and Software Engineering*, pages 173–177, 2023.
199. X.-Z. Zhang, W.-L. Zheng, B.-L. Lu, Liu D, Xie S, Li Y, El-Alfy E.M, and Zhao D. EEG-Based Sleep Quality Evaluation with Deep Transfer Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10637:543–552, 2017.
200. Hongle Wu, Takafumi Kato, Masayuki Numao, and Ken-Ichi Fukui. Statistical sleep pattern modelling for sleep quality assessment based on sound events. *Health information science and systems*, 5(1):11, December 2017.
201. Ji-Hyeok Park and Jae-Dong Lee. A Customized Deep Sleep Recommender System Using Hybrid Deep Learning. *Sensors (Basel, Switzerland)*, 23(15), July 2023.
202. C. Wang, F.W. Usher, S.A. Alvarez, C. Ruiz, and M. Moonis. Clustering of Human Sleep Recordings Using a Quantile Representation of Stage Bout Durations. *Communications in Computer and Information Science*, 357:369–384, 2013.
203. F.W. Usher, C. Wang, S.A. Alvarez, C. Ruiz, and M. Moonis. Machine learning of human sleep patterns based on stage bout durations. *HEALTHINF 2012 - Proceedings of the International Conference on Health Informatics*, pages 71–80, 2012.
204. C. Wang, S.A. Alvarez, C. Ruiz, M. Moonis, Gamboa H, and Fred A. Modeling and clustering of human sleep time series using dynamic time warping: Sequential and distributed implementations. *Communications in Computer and Information Science*, 690:276–294, 2017.
205. T.M. Bajkowski, N. Marchal, J. Saied-Walker, P. Gupta, J.M. Keller, M. Skubic, and G.J. Scott. Cohort Discovery from Bed Sensor Data with Fuzzy Evidence Accumulation Clustering. *IEEE International Conference on Fuzzy Systems*, 2023.
206. A. Khasawneh, S.A. Alvarez, C. Ruiz, S. Misra, and M. Moonis. Discovery of sleep composition types using expectation-maximization. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, pages 26–31, 2010.
207. A. Khasawneh, S.A. Alvarez, C. Ruiz, S. Misra, and M. Moonis. Similarity grouping of human sleep recordings using EEG and ECG. *Communications in Computer and Information Science*, 273:380–394, 2011.
208. Meredith L. Wallace, Daniel J. Buysse, Anne Germain, Martica H. Hall, and Satish Iyengar. Variable Selection for Skewed Model-Based Clustering: Application to the Identification of Novel Sleep Phenotypes. *Journal of the American Statistical Association*, 113(521):95–110, 2018.
209. D.L. Piza, A. Schulze-Bonhage, T. Stieglitz, J. Jacobs, and M. Dumpelmann. Depuration, augmentation and balancing of training data for supervised learning based detectors of EEG patterns. *International IEEE/EMBS Conference on Neural Engineering, NER*, pages 497–500, 2017.
210. Zheng Chen, Pei Gao, Ming Huang, Naoaki Ono, M. D. Altaf-Ul-Amin, and Shigehiko Kanaya. Feasibility Analysis of Symbolic Representation for Single-Channel EEG-Based Sleep Stages. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2021: 5928–5931, November 2021.
211. R. Ranjan, R. Arya, S.L. Fernandes, E. Sravya, and V. Jain. A fuzzy neural network approach for automatic K-complex detection in sleep EEG signal. *Pattern Recognition Letters*, 115:74–83, 2018.
212. Evangelia I. Zacharaki, Evangelia Pippa, Andreas Koupparis, Vasileios Kokkinos, George K. Kostopoulos, and Vasileios Megalooikonomou. One-class classification of temporal EEG patterns for K-complex extraction. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2013:5801–5804, 2013.
213. G. Safont, A. Salazar, L. Vergara, E. Gomez, and V. Villanueva. Mixtures of independent component analyzers for microarousal detection. *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI 2014*, pages 752–755, 2014.
214. G. Safont, A. Salazar, L. Vergara, and Herencsar N. Unsupervised learning of non-Gaussian mixtures with temporal dependencies. *2017 40th International Conference on Telecommunications and Signal Processing, TSP 2017*, 2017:399–402, 2017.
215. Fábio Mendonça, Sheikh Shanawaz Mostafa, Fernando Morgado-Dias, Antonio G. Ravelo-García, and Thomas Penzel. Sleep quality of subjects with and without sleep-disordered breathing based on the cyclic alternating pattern rate estimation from single-lead ECG. *Physiological measurement*, 40(10):105009, November

2019.

216. Chanakya Reddy Patti, Thomas Penzel, and Dean Cvetkovic. Automated sleep spindle detection using IIR filters and a Gaussian Mixture Model. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2015:610–613, August 2015.
217. C.R. Patti, R. Chaparro-Vargas, and D. Cvetkovic. Automated Sleep Spindle detection using novel EEG features and mixture models. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, pages 2221–2224, 2014.
218. M. He, P. Das, G. Hotan, P.L. Purdon, and Matthews M.B. Automatic Segmentation of Sleep Spindles: A Variational Switching State-Space Approach. *Conference Record - Asilomar Conference on Signals, Systems and Computers*, 2022:1301–1305, 2022.
219. C. Chen, X. Zhu, A.N. Belkacem, L. Lu, L. Hao, J. You, D. Shin, W. Tan, Z. Huang, D. Ming, and Wang Y. Automatic Sleep Spindle Detection and Analysis in Patients with Sleep Disorders. *Communications in Computer and Information Science*, 1369:113–124, 2021.
220. Christian O'Reilly, Jonathan Godbout, Julie Carrier, and Jean-Marc Lina. Combining time-frequency and spatial information for the detection of sleep spindles. *Frontiers in human neuroscience*, 9:70, 2015.
221. Chao Chen, Jiayuan Meng, Abdelkader Nasreddine Belkacem, Lin Lu, Fengyue Liu, Weibo Yi, Penghai Li, Jun Liang, Zhaoyang Huang, and Dong Ming. Hierarchical fusion detection algorithm for sleep spindle detection. *Frontiers in neuroscience*, 17:1105696, 2023.
222. Erricos M. Ventouras, Periklis Y. Ktonas, Hara Tsekou, Thomas Paparrigopoulos, Ioannis Kalatzis, and Constantin R. Soldatos. Independent component analysis for source localization of EEG sleep spindle components. *Computational intelligence and neuroscience*, 2010: 329436, 2010.
223. Chanakya Reddy Patti, Thomas Penzel, and Dean Cvetkovic. Sleep spindle detection using multivariate Gaussian mixture models. *Journal of sleep research*, 27 (4):e12614, August 2018.
224. E.M. Ventouras, P.Y. Ktonas, H. Tsekou, T. Paparrigopoulos, I. Kalatzis, and C.R. Soldatos. Slow and fast EEG sleep spindle component extraction using Independent Component Analysis. *8th IEEE International Conference on BioInformatics and BioEngineering, BIBE 2008*, 2008.
225. Y. Gu, C. Zhang, Y. Wang, Z. Liu, Y. Ji, and J. Li. A Contactless and Fine-Grained Sleep Monitoring System Leveraging WiFi Channel Response. *IEEE International Conference on Communications*, 2019, 2019.
226. A. Heinrich, X. Zhao, and G. De Haan. Multi-distance motion vector clustering algorithm for video-based sleep analysis. *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services, Healthcom 2013*, pages 223–227, 2013.
227. H. Wu, T. Kato, T. Yamada, M. Numao, K.-I. Fukui, Ali M, Fujita H, Sasaki J, Kurematsu M, and Selamat A. Sleep pattern discovery via visualizing cluster dynamics of sound data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9799:460–471, 2016.
228. Shane A Landry, Caroline Beatty, Luke DJ Thomson, Ai-Ming Wong, Bradley A Edwards, Garun S Hamilton, and Simon A Joosten. A review of supine position related obstructive sleep apnea: classification, epidemiology, pathogenesis and treatment. *Sleep Medicine Reviews*, page 101847, 2023.
229. Lourdes M DelRosso, Rosalia Silvestri, and Raffaele Ferri. Restless sleep disorder. *Sleep medicine clinics*, 16(2):381–387, 2021.
230. D.V. Dhongade, T.V.K.H. Rao, Prabakar S, Porkumaran K, and Jaganathan S. Classification of sleep disorders based on EEG signals by using feature extraction techniques with KNN classifier. *IEEE International Conference on Innovations in Green Energy and Healthcare Technologies - 2017, IGEHT 2017*, 2017.
231. T. Wongsirichot, A. Hanskunatai, Bevilacqua V, Huang D.-S, and Premaratne P. A comparative investigation of PSG signal patterns to classify sleep disorders using machine learning techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9225: 510–521, 2015.
232. Scott A. Bruce. Adaptive Clustering and Feature Selection for Categorical Time Series Using Interpretable Frequency-Domain Features. *Statistics and its interface*, 16(2):319–335, 2023.
233. P. Moeynoi and Y. Kitjaidure. Canonical correlation analysis for dimensionality reduction of sleep apnea features based on ECG single lead. *BMEiCON 2016 - 9th Biomedical Engineering International Conference*, 2017.
234. Shuaicong Hu, Ya'nan Wang, Jian Liu, Cuiwei Yang, Aiguo Wang, Kuanzheng Li, and Wenxin Liu. Semi-Supervised Learning for Low-Cost Personalized Obstructive Sleep Apnea Detection Using Unsupervised Deep Learning and Single-Lead Electrocardiogram. *IEEE journal of biomedical and health informatics*, 27(11): 5281–5292, November 2023.
235. K. Feng, H. Qin, S. Wu, W. Pan, and G. Liu. A Sleep Apnea Detection Method Based on Unsupervised Feature Learning and Single-Lead Electrocardiogram. *IEEE Transactions on Instrumentation and Measurement*, 70, 2021.
236. A.G. Ravelo-García, J.L. Navarro-Mesa, M.J. Murillo-Díaz, and J.G. Juliá-Serdá. Application of RR series and oximetry to a statistical classifier for the detection of sleep apnoea/Hypopnoea. *Computers in Cardiology*, 31:305–308, 2004.
237. D.Y.Y. Sim, A.I. Ismail, and C.S. Teh. Effective k-Means Clustering in Greedy Prepruned Tree-based Classification for Obstructive Sleep Apnea. *International Journal of Electrical and Electronic Engineering and Telecommunications*, 11(3):242–248, 2022.
238. T. Al-Ani, C.K. Karmakar, A.H. Khandoker, and M. Palaniswami. Automatic recognition of obstructive sleep apnoea syndrome using power spectral analysis of electrocardiogram and hidden markov models. *ISSNIP 2008 - Proceedings of the 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pages 285–290, 2008.
239. D. Novák, D. Cuesta-Frau, T. Al ani, M. Aboy, P. Mico, and L. Lhotská. Speech recognition methods applied to biomedical signals processing. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE*

Engineering in Medicine and Biology Society. Annual Conference, 2006:118–121, 2004.

240. K. Feng and G. Liu. Obstructive sleep apnea detection based on unsupervised feature learning and hidden markov model. *3rd International Conference on Biological Information and Biomedical Engineering, BIBE 2019*, pages 110–113, 2019.
241. Javad Ostadieh and Mehdi Chehel Amirani. Introducing the Hybrid "K-means, RLS" Learning for the RBF Network in Obstructive Apnea Disease Detection using Dual-tree Complex Wavelet Transform Based Features. *Journal of electrical bioimpedance*, 11(1):4–11, January 2020.
242. S. Boudaoud, C. Heneghan, H. Rix, O. Meste, and C. O'Brien. P-wave shape changes observed in the surface electrocardiogram of subjects with obstructive sleep apnoea. *Computers in Cardiology*, 32:359–362, 2005.
243. U.M. Boppana, P. Ranjana, K. Dhivyapriya, and D. Nagarajan. Analyzingobstructive sleep apnea(OSA)using machine perception and wavelet taransforms. *International Journal of Engineering and Advanced Technology*, 8(4):78–85, 2019.
244. Daniel Alvarez, Roberto Hornero, J. Víctor Marcos, Félix del Campo, and Miguel López. Obstructive sleep apnea detection using clustering classification of nonlinear features from nocturnal oximetry. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2007: 1937–1940, 2007.
245. H.J. Robertson, J.J. Soraghan, C. Idzikowski, and B.A. Conway. EMD and PCA for the prediction of sleep apnoea: A comparative study. *ISSPIT 2007 - 2007 IEEE International Symposium on Signal Processing and Information Technology*, pages 419–424, 2007.
246. Javad Ostadieh, Mehdi Chehel Amirani, and Morteza Valizadeh. Enhancing Obstructive Apnea Disease Detection Using Dual-Tree Complex Wavelet Transform-Based Features and the Hybrid "K-Means, Recursive Least-Squares" Learning for the Radial Basis Function Network. *Journal of medical signals and sensors*, 10(4): 219–227, October 2020.
247. P. Kumar Tyagi and D. Agrawal. Automatic detection of sleep apnea from single-lead ECG signal using enhanced-deep belief network model. *Biomedical Signal Processing and Control*, 80, 2023.
248. C.B. Kumar, A.K. Mondal, M. Bhatia, B.K. Panigrahi, and T.K. Gandhi. Self-Supervised Representation Learning-Based OSA Detection Method Using Single-Channel ECG Signals. *IEEE Transactions on Instrumentation and Measurement*, 72, 2023.
249. K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu. A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ECG signal. *Neurocomputing*, 294:94–101, 2018.
250. M. Zubair, U.K.N. M, R.K. Tripathy, M. Alhartomi, S. Alzahrani, and S.R. Ahamed. Detection of Sleep Apnea from ECG Signals using Sliding Singular Spectrum based Sub-Pattern Principal Component Analysis. *IEEE Transactions on Artificial Intelligence*, pages 1–10, 2023.
251. L. M. Sepúlveda-Cano, E. Gil, P. Laguna, and G. Castellanos-Dominguez. Sleep apnoea detection in children using PPG envelope-based dynamic features. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2011:1483–1486, 2011.
252. H. Sharma and K.K. Sharma. Sleep apnea detection from ECG using variational mode decomposition. *Biomedical Physics and Engineering Express*, 6(1), 2020.
253. 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. *TIC-STH'09: 2009 IEEE Toronto International Conference - Science and Technology for Humanity*, pages 279–284, 2009.
254. S. Boudaoud, H. Rix, O. Meste, C. Heneghan, and C. O'Brien. Corrected integral shape averaging applied to obstructive sleep apnea detection from the electrocardiogram. *Eurasip Journal on Advances in Signal Processing*, 2007, 2007.
255. P. Temrat, Y. Jiraraksopakun, K. Weaasae, and A. Bhatranand. Suitable feature selection for OSA classification based on snoring sounds. *ECTI-CON 2018 - 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pages 396–399, 2018.
256. Y. Zigel, A. Tarasiuk, and E. Goldshtein. Analysis of speech signals among obstructive sleep apnea patients. *IEEE Convention of Electrical and Electronics Engineers in Israel, Proceedings*, pages 760–764, 2008.
257. J.L. Blanco, R. Fernández, D. Pardo, Á. Sigüenza, L.A. Hernández, and J. Alcázar. Analyzing GMMs to characterize resonance anomalies in speakers suffering from apnoea. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1459–1462, 2009.
258. J.L. Blanco, R. Fernández, D. Torre, F.J. Caminero, and E. López. Analyzing training dependencies and posterior fusion in discriminant classification of apnea patients based on sustained and connected speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3033–3036, 2011.
259. R.F. Pozo, J.L.B. Murillo, L.H. Gmez, E.L. Gonzalo, J.A. Ramírez, and D.T. Toledano. Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques. *Eurasip Journal on Advances in Signal Processing*, 2009, 2009.
260. O. Elisha, A. Tarasiuk, Y. Zigel, and Manfredi C. Detection of obstructive sleep apnea using speech signal analysis. *Models and Analysis of Vocal Emissions for Biomedical Applications - 7th International Workshop, MAVEBA 2011*, pages 13–16, 2011.
261. R. Fernández, J.L. Blanco, D. Díaz, L.A. Hernández, E. López, J. Alcázar, Fred A, Gamboa H, Lisbon Technical University of Lisbon, Institute of Telecommunications, Filipe J, and Informatics Setubal Polytechnic Institute of Setubal, Departament of Systems and. Early detection of severe apnoea through voice analysis and automatic speaker recognition techniques. *Communications in Computer and Information Science*, 52:245–257, 2010.
262. Wan-Ju Cheng, Eysteinn Finnsson, Eydís Arnardóttir, Jón S Ágústsson, Scott A Sands, and Liang-Wen Hang. Relationship between symptom profiles and endotypes among patients with obstructive sleep apnea: A latent class analysis. *Annals of the American Thoracic Society*,

20(9):1337–1344, 2023.

263. H. Alshaer, M. Garcia, M.H. Radfar, G.R. Fernie, and T.D. Bradley. Detection of upper airway narrowing via classification of LPC coefficients: Implications for obstructive sleep apnea diagnosis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 681–684, 2011.
264. Y. Wu, Z. Zhao, K. Qian, Z. Xu, and H. Xu. Analysis of long duration snore related signals based on formant features. *Proceedings - 2013 International Conference on Information Technology and Applications, ITA 2013*, pages 91–95, 2013.
265. Arun Sebastian, Peter A. Cistulli, Gary Cohen, and Philip de Chazal. Unsupervised Approach for the Identification of the Predominant Site of Upper Airway Collapse in Obstructive Sleep Apnoea Patients Using Snore Signals. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2021:160–163, November 2021.
266. J.F. Rodrigues, S. Bailly, J.-L. Pepin, L. Goeuriot, G. Spadon, and S. Amer-Yahia. CPAP Adherence Assessment via Gaussian Mixture Modeling of Telemonitored Apnea Therapy. *Applied Sciences (Switzerland)*, 12(15), 2022.
267. H.E. Romero, N. Ma, G.J. Brown, A.V. Beeston, and M. Hasan. Deep Learning Features for Robust Detection of Acoustic Events in Sleep-disordered Breathing. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019:810–814, 2019.
268. Valeria Mordoh and Yaniv Zigel. Audio source separation to reduce sleeping partner sounds: a simulation study. *Physiological measurement*, 42(6), June 2021.
269. T. Wongsirichot, N. Iad-ua, J. Wibulkit, Burduk R, Jackowski K, Kurzyński M, Woźniak M, and Żołnierek A. A snoring sound analysis application using k-mean clustering method on mobile devices. *Advances in Intelligent Systems and Computing*, 403:789–796, 2016.
270. M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller. A bag-of-audio-words approach for snore sounds' excitation localisation. *Speech Communication - 12. ITG-Fachtagung Sprachkommunikation*, pages 230–234, 2016.
271. Ali Azarbarzin and Zahra M. K. Moussavi. Automatic and unsupervised snore sound extraction from respiratory sound signals. *IEEE transactions on bio-medical engineering*, 58(5):1156–1162, May 2011.
272. E. Dafna, A. Tarasiuk, Y. Zigel, and Manfredi C. Automatic detection of snoring events using Gaussian mixture models. *Models and Analysis of Vocal Emissions for Biomedical Applications - 7th International Workshop, MAVEBA 2011*, pages 17–20, 2011.
273. Azadeh Yadollahi and Zahra Moussavi. Formant analysis of breath and snore sounds. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2009: 2563–2566, 2009.
274. F. Vrins, J. Deswert, D. Bouvy, V. Bouillon, J.A. Lee, C. Eugène, and M. Verleysen. On the extraction of the snore acoustic signal by independent component analysis. *Proceedings of the IASTED International Conference on Biomedical Engineering*, pages 326–331, 2004.
275. Ali Azarbarzin and Zahra Moussavi. Unsupervised classification of respiratory sound signal into snore/no-snore classes. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2010:3666–3669, 2010.
276. C.F. Bublitz, A.C. Ribeiro-Teixeira, T.A. Pianoschi, J. Rochol, and C.B. Both. Unsupervised segmentation and classification of snoring events for Mobile Health. *2017 IEEE Global Communications Conference, GLOBECOM 2017 - Proceedings*, 2018:1–6, 2017.
277. G. Ma, B. Xue, H. Hong, X. Zhu, and Z. Wang. Unsupervised snore detection from respiratory sound signals. *International Conference on Digital Signal Processing, DSP*, 2015:417–421, 2015.
278. Sungkyu Park, Sang Won Lee, Sungwon Han, and Meeyoung Cha. Clustering Insomnia Patterns by Data From Wearable Devices: Algorithm Development and Validation Study. *JMIR mHealth and uHealth*, 7(12): e14473, December 2019.
279. S.T.-B. Hamida, B. Ahmed, and T. Penzel. A novel insomnia identification method based on Hjorth parameters. *2015 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2015*, pages 548–552, 2016.
280. D.L. Frederic, A. Rouijel, H. Elghazi, Mohamed B.A, Abdelhakim B.A, and Mazri T. Insomnia eeg signal preprocessing using ICA algorithms. *ACM International Conference Proceeding Series*, 2021.
281. S. Tripathi, P. Mattioli, C. Liguori, A. Chiaravalloti, D. Arnaldi, and L. Giancardo. Brain Hemisphere Dissimilarity, a Self-Supervised Learning Approach for alpha-synucleinopathies prediction with FDG PET. *Proceedings. IEEE International Symposium on Biomedical Imaging*, 2023, April 2023.
282. Henriette Koch, Julie A. E. Christensen, Rune Frandsen, Lars Arvastson, Soren R. Christensen, Helge B. D. Sorensen, and Poul Jennum. Classification of iRBD and Parkinson's patients using a general data-driven sleep staging model built on EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2013: 4275–4278, 2013.
283. Mehrnaz Shokrollahi and Sridhar Krishnan. Sleep EMG analysis using sparse signal representation and classification. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2012:3480–3483, 2012.
284. Mauro Manconi, Diego Garcia-Borreguero, Barbara Schormair, Aleksandar Videnovic, Klaus Berger, Raffaele Ferri, and Yves Dauvilliers. Restless legs syndrome. *Nature reviews Disease primers*, 7(1):80, 2021.
285. Magdolna Hornyak, Bernd Feige, Dieter Riemann, and Ulrich Voderholzer. Periodic leg movements in sleep and periodic limb movement disorder: prevalence, clinical significance and treatment. *Sleep medicine reviews*, 10(3): 169–177, 2006.
286. A.M. Adami, A. Adami, C.M. Singer, T.L. Hayes, and M. Pavel. A System for unobtrusive monitoring of mobility in bed. *Proceedings of the 11th IEEE International Conference on Computational Science and Engineering, CSE Workshops 2008*, pages 13–18, 2008.

287. N. Cooray, Z. Li, J. Wang, C. Lo, M. Arvaneh, M. Symmonds, M. Hu, M. De Vos, and L.S. Mihaylova. Automated Movement Detection with Dirichlet Process Mixture Models and Electromyography. *2022 25th International Conference on Information Fusion, FUSION 2022*, 2022.
288. Jacqueline A. Fairley, George Georgoulas, Otis L. Smart, George Dimakopoulos, Petros Karvelis, Chrysostomos D. Stylios, David B. Rye, and Donald L. Bliwise. Wavelet analysis for detection of phasic electromyographic activity in sleep: influence of mother wavelet and dimensionality reduction. *Computers in biology and medicine*, 48:77–84, May 2014.
289. P. Staroniewicz. Effect of Sleepiness in the Voice on Speaker Recognition Performance. *Vibrations in Physical Systems*, 32(2), 2021.
290. Shahin Amiriparian, Pawel Winokurow, Vincent Karas, Sandra Ottl, Maurice Gerczuk, and Björn Schuller. Unsupervised representation learning with attention and sequence to sequence autoencoders to predict sleepiness from speech. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 11–17, 2020.
291. Fabienne Porée, Amar Kachenoura, Hervé Gauvrit, Catherine Morvan, Guy Carrault, and Lotfi Senhadji. Blind source separation for ambulatory sleep recording. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 10(2):293–301, April 2006.
292. Ciprian M. Crainiceanu, Brian S. Caffo, Sheng Luo, Vadim M. Zipunnikov, and Naresh M. Punjabi. Population Value Decomposition, a Framework for the Analysis of Image Populations. *Journal of the American Statistical Association*, 106(495), 2011.
293. Q.T.N. Nguyen, T. Le, Q.T.T. Vu, K.D. Bui, H.T. Ngo, Van Toi V, Nguyen T.-H, Long V.B, and Huong H.T.T. An Algorithm for Removing Artifacts in Polysomnography Signals. *IFMBE Proceedings*, 85:1017–1031, 2022.
294. R.N. Sekkal, F. Bereksi-Reguig, N. Dib, D. Ruiz-Fernandez, Rojas I, Valenzuela O, Rojas F, Herrera L.J, and Ortuño F. An Approach to Detecting and Eliminating Artifacts from the Sleep EEG Signals. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12108:155–160, 2020.
295. M. Zima, P. Tichavsk, K. Paul, and V. Krajča. Robust removal of short-duration artifacts in long neonatal EEG recordings using wavelet-enhanced ICA and adaptive combining of tentative reconstructions. *Physiological Measurement*, 33(8):N39–N49, 2012.
296. Maite Crespo-Garcia, Mercedes Atienza, and Jose L. Cantero. Muscle artifact removal from human sleep EEG by using independent component analysis. *Annals of biomedical engineering*, 36(3):467–475, March 2008.
297. A. Kachenoura, H. Gauvrit, and L. Senhadji. Extraction and separation of eyes movements and the muscular tonus from a restricted number of electrodes using the Independent Component Analysis. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 3:2359–2362, 2003.
298. S Devuyst, T Dutoit, P Stenuit, M Kerkhofs, and E Stanus. Cancelling ECG Artifacts in EEG Using a Modified Independent Component Analysis Approach. *EURASIP JOURNAL ON ADVANCES IN SIGNAL PROCESSING*, 2008.
299. S. Romero, M.A. Mañanas, S. Clos, S. Gimenez, and M.J. Barbanoj. Reduction of EEG Artifacts by ICA in Different Sleep Stages. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 3:2675–2678, 2003.
300. Yves Leclercq, Evelyne Balteau, Thanh Dang-Vu, Manuel Schabus, André Luxen, Pierre Maquet, and Christophe Phillips. Rejection of pulse related artefact (PRA) from continuous electroencephalographic (EEG) time series recorded during functional magnetic resonance imaging (fMRI) using constraint independent component analysis (cICA). *NeuroImage*, 44(3):679–691, February 2009.
301. R. Sameni, M.B. Shamsollahi, and L. Senhadji. Processing polysomnographic signals, using independent component analysis approaches. *Proceedings of the IASTED International Conference on Biomedical Engineering*, pages 193–196, 2004.
302. Richard Somervail, Jacinthe Cataldi, Aurélie M. Stephan, Francesca Siclari, and Gian Domenico Iannetti. Dusk2Dawn: an EEGLAB plugin for automatic cleaning of whole-night sleep electroencephalogram using Artifact Subspace Reconstruction. *Sleep*, 46(12):zsad208, December 2023.
303. Jacqueline Fairley, Ashley N. Johnson, George Georgoulas, and George Vachtsevanos. Automated polysomnogram artifact compensation using the generalized singular value decomposition algorithm. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2010: 5097–5100, 2010.
304. X. Dong, J. Zhang, G. Wang, Y. Xia, Lin Z, Wang L, Tan T, Yang J, Shi G, Zheng N, Chen X, and Zhang Y. DAEimp: Denoising autoencoder-based imputation of sleep heart health study for identification of cardiovascular diseases. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11857:517–527, 2019.
305. H. Rajaguru, R. Karthikamani, Harikumar R, Babu C.G, Poongodi C, and Deepa D. Performance Analysis of the Classifier in the Classification of Normal-Sleep and Seizure from EEG Signal. *Proceedings - 2nd International Conference on Smart Technologies, Communication and Robotics 2022, STCR 2022*, 2022.
306. Xiaoran Sun, Li Ding, Yujun Song, Jianxin Peng, Lijuan Song, and Xiaowen Zhang. Automatic identifying OSAHS patients and simple snorers based on Gaussian mixture models. *Physiological measurement*, 44(4), May 2023.
307. Joonas Paalasmaa, Hannu Toivonen, and Markku Partinen. Adaptive Heartbeat Modeling for Beat-to-Beat Heart Rate Measurement in Ballistocardiograms. *IEEE journal of biomedical and health informatics*, 19(6): 1945–1952, November 2015.
308. W. Wong, V. Noreika, L. Móró, A. Revonsuo, J. Windt, K. Valli, and N. Tsuchiya. The Dream Catcher experiment: Blinded analyses failed to detect markers of dreaming consciousness in EEG spectral power. *Neuroscience of Consciousness*, 2020(1), 2020.
309. John A. Bentrup. Reliable signal classification using quality assurance and clustering - an application to sleep stage scoring. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 3:2717–2722, 1994.

310. W. Lee, G. Kim, J. Yu, and Y. Kim. Model Interpretation Considering Both Time and Frequency Axes Given Time Series Data. *Applied Sciences (Switzerland)*, 12(24), 2022.
311. Li-Chen Shi and Bao-Liang Lu. Dynamic clustering for vigilance analysis based on EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2008: 54–57, 2008.
312. L.-C. Shi, H. Yu, and B.-L. Lu. Semi-supervised clustering for vigilance analysis based on EEG. *IEEE International Conference on Neural Networks - Conference Proceedings*, pages 1518–1523, 2007.
313. G.K. Rajini, S. Naseera, S. Pandey, and A. Bhuvaneswaran. Classification of EEG analysis for sleep detection using wavelet transform and neural network. *Lecture Notes in Networks and Systems*, 11: 523–533, 2018.
314. Y. Yin, Y. Zhu, S. Xiong, and J. Zhang. Drowsiness detection from EEG spectrum analysis. *Lecture Notes in Electrical Engineering*, 133:753–759, 2011.
315. N. Gurudath, H. Bryan Riley, and Shakshuki E.M. Drowsy driving detection by EEG analysis using Wavelet Transform and K-means clustering. *Procedia Computer Science*, 34:400–409, 2014.
316. Mu Li, Jia-Wei Fu, and Bao-Liang Lu. Estimating vigilance in driving simulation using probabilistic PCA. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2008:5000–5003, 2008.
317. D. Sommer, M. Golz, Hornik K, Augasse 2-6 Wien 1090 Wirtschaftsuniversität Wien, Institut für Statistik, Dorffner G, Dept. of Mecidal Cybernetics University of Vienna, Vienna 1010 Artificial Intelligence, Freyung 6/2, Bischof H, Pattern Recognition Technical University of Vienna, Institute for Computer Aided Automation, and Vienna 1040 Image Processing Group, Favoritenstr. 9/1832. Clustering of EEG-Segments using hierarchical agglomerative methods and self-organizing maps. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2130:642–649, 2001.
318. Seyed Mohammad Reza Noori and Mohammad Mikaeili. Driving Drowsiness Detection Using Fusion of Electroencephalography, Electrooculography, and Driving Quality Signals. *Journal of medical signals and sensors*, 6(1):39–46, January 2016.
319. A. Dutta, S. Kour, and S. Taran. Automatic drowsiness detection using electroencephalogram signal. *Electronics Letters*, 56(25):1383–1386, 2020.
320. O. Maftukhaturrizqoh, N. Nuryani, and D. Darmanto. Drowsiness detection using radial basis function network with electrocardiographic RR interval statistical feature. *Journal of Physics: Conference Series*, 1153(1), 2019.
321. Sheng-Hsiou Hsu and Tzyy-Ping Jung. Monitoring alert and drowsy states by modeling EEG source nonstationarity. *Journal of neural engineering*, 14(5): 056012, October 2017.
322. W.Y. Leong and D.P. Mandic. Cascaded approach for microsleep data extraction. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 2045–2048, 2008.
323. Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.
324. Lane Fitzsimmons, Maya Dewan, and Judith W Dexheimer. Diversity in machine learning: A systematic review of text-based diagnostic applications. *Applied Clinical Informatics*, 13(03):569–582, 2022.
325. Jesmin Nahar, Tasadduq Imam, Kevin S Tickle, and Yi-Ping Phoebe Chen. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert systems with applications*, 40(4):1086–1093, 2013.
326. Murat Karabatak and M Cevdet Ince. An expert system for detection of breast cancer based on association rules and neural network. *Expert systems with Applications*, 36(2):3465–3469, 2009.
327. Hao Guan, Li Wang, and Mingxia Liu. Multi-source domain adaptation via optimal transport for brain dementia identification. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1514–1517. IEEE, 2021.
328. C.-F. Chao, J.-A. Jiang, M.-J. Chiu, and R.-G. Lee. Automated long-term polysomnography analysis with wavelet processing and adaptive fuzzy clustering. *Biomedical Engineering - Applications, Basis and Communications*, 18(3):119–123, 2006.
329. M. Dursun, S. Gunes, S. Ozsen, and S. Yosunkaya. Comparison of artificial immune clustering with fuzzy C-means clustering in the sleep stage classification problem. *INISTA 2012 - International Symposium on INnovations in Intelligent SysTems and Applications*, 2012.
330. K.S. Prabhudesai, L.M. Collins, and B.O. Mainsah. Automated feature learning using deep convolutional auto-encoder neural network for clustering electroencephalograms into sleep stages. *International IEEE/EMBS Conference on Neural Engineering, NER*, 2019:937–940, 2019.
331. C. Wang, S.A. Alvarez, C. Ruiz, and M. Moonis. Semi-markov modeling-clustering of human sleep with efficient initialization and stopping. *BIOSIGNALS 2014 - 7th Int. Conference on Bio-Inspired Systems and Signal Processing, Proceedings; Part of 7th Int. Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2014*, pages 61–68, 2014.
332. Junming Zhang and Yan Wu. Complex-valued unsupervised convolutional neural networks for sleep stage classification. *Computer methods and programs in biomedicine*, 164:181–191, October 2018.
333. Y. Yu, B. Wang, J. Jin, X. Wang, Li Q, and Wang L. Automatic Sleep Stage Classification by a Density - Distance- Based K - means Clustering Algorithm with Amendments. *Proceedings - 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2019*, 2019.
334. X. Zheng, X. Yin, X. Shao, Y. Li, and X. Yu. Collaborative Sleep Electroencephalogram Data Analysis Based on Improved Empirical Mode Decomposition and Clustering Algorithm. *Complexity*, 2020, 2020.
335. M. Obayya and F.E.Z. Abou-Chadi. Automatic classification of sleep stages using EEG records based on Fuzzy c-means (FCM) algorithm. *National Radio Science Conference, NRSC, Proceedings*, pages 265–272, 2014.

336. S. Raiesdana, S. Esmaielzadehha, and A. Banaie. Comparison of different clustering algorithms applied to nonliner features for sleep stages discrimination. *13th Iranian Conference on Fuzzy Systems, IFSC 2013*, 2013.
337. I.N. Yulita, R. Rosadi, S. Purwani, and Amien M. Sleep stage classification using fuzzy long short-term memory. *Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology, CAIPT 2017*, 2018:1–5, 2017.
338. I.N. Yulita, M.I. Fanany, A.M. Arymuthy, Budiharto W, Suryani D, Wulandhari L.A, Chowanda A, Gunawan A.A.S, Hanafiah N, Ham H, and Meiliana null. Bi-directional Long Short-Term Memory using Quantized data of Deep Belief Networks for Sleep Stage Classification. *Procedia Computer Science*, 116:530–538, 2017.
339. C.-F. Chao, J.-A. Jiang, M.-J. Chiu, and R.-G. Lee. Wavelet-based processing and adaptive fuzzy clustering for automated long-term polysomnography analysis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2:II1176–II1179, 2006.
340. Efendi Nasibov, Murat Ozgören, Gözde Ulutagay, Adile Oniz, and Sibel Kocaaslan. On the analysis of BIS stage epochs via fuzzy clustering. *Biomedizinische Technik. Biomedical engineering*, 55(3):147–153, June 2010.
341. S. Rajalakshmi, R. Venkatesan, Das S, Krishnan S, Corchado Rodriguez J.M, Wozniak M, Al-Jumeily D, and Thampi S.M. Exploring cepstral coefficient based sleep stage scoring method for single-channel EEG signal using machine learning technique. *Advances in Intelligent Systems and Computing*, 678:24–37, 2018.
342. M. Koivuluoma, A. Värri, and A. Flexer. Modelling sleep with Gaussian mixture model based on eye movements and delta-activity. *European Signal Processing Conference*, 2015, 2000.
343. G. Garcia-Molina, M. Bellesi, S. Pastoor, S. Pfundtner, B. Riedner, and G. Tononi. Online single EEG channel based automatic sleep staging. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8020:333–342, 2013.
344. U. Rajendra Acharya, Eric Chern-Pin Chua, Kuang Chua Chua, Lim Choo Min, and Toshiyo Tamura. Analysis and automatic identification of sleep stages using higher order spectra. *International journal of neural systems*, 20(6): 509–521, December 2010.
345. S. Rajalakshmi, R. Prakash, R. Venkatesan, and A.B. Ganesh. Sleep stage scoring based on single-channel EEG using machine learning technique. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*, pages 916–921, 2018.
346. Y. Guo, H.X. Mao, J. Yin, and Z.-H. Mao. Gaussian transformation enhanced semi-supervised learning for sleep stage classification. *Journal of Big Data*, 10(1), 2023.
347. Tarek Lajnef, Sahbi Chaibi, Perrine Ruby, Pierre-Emmanuel Aguera, Jean-Baptiste Eichenlaub, Mounir Samet, Abdennaceur Kachouri, and Karim Jerbi. Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. *Journal of neuroscience methods*, 250:94–105, 2015.
348. V. Gerla, V. Djordjevic, L. Lhotska, and V. Krajca. System approach to complex signal processing task. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5717:579–586, 2009.
349. V Gerla, V Kremen, M Macas, D Dudysova, A Mladek, P Sos, and L Lhotska. Iterative expert-in-the-loop classification of sleep PSG recordings using a hierarchical clustering. *Journal of Neuroscience Methods*, 317:61–70, April 2019.
350. V. Gerla, M. Murgas, A. Mladek, E. Saifutdinova, M. Macas, L. Lhotska, Maglaveras N, Chouvarda I, Maglaveras N, and de Carvalho P. Hybrid hierarchical clustering algorithm used for large datasets: A pilot study on long-term sleep data. *IFMBE Proceedings*, 66:3–7, 2018.
351. A. Kazemi, M.J. McKeown, and M.S. Mirian. Sleep staging using semi-unsupervised clustering of EEG: Application to REM sleep behavior disorder. *Biomedical Signal Processing and Control*, 75, 2022.
352. J.L. Rodríguez-Sotelo, A. Osorio-Forero, A. Jiménez-Rodríguez, D. Cuesta-Frau, E. Cirugeda-Roldán, and D. Peluffo. Automatic sleep stages classification using EEG entropy features and unsupervised pattern analysis techniques. *Entropy*, 16(12):6573–6589, 2014.
353. J.L. Rodríguez-Sotelo, A. Osorio-Forero, A. Jiménez-Rodríguez, F. Restrepo-De-Mejía, D.H. Peluffo-Ordoñez, J. Serrano, Adeli H, Ferrandez Vicente J.M, Toledo Moreo J, Alvarez-Sanchez J.R, and de la Paz Lopez F. Sleep stages clustering using time and spectral features of EEG signals an unsupervised approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10337:444–455, 2017.
354. X. Shao, B. Hu, X. Zheng, Zeng Y, Xu B, Martone M, He Y, Peng H, Luo Q, and Kotaleski J.H. A Study on Automatic Sleep Stage Classification Based on Clustering Algorithm. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10654:139–148, 2017.
355. Xingjun Wang and Ziyao Xu. Automatic Sleep Staging based on Curriculum Learning Approach. pages 1–6, 2019.
356. Nicolas Decat, Jasmine Walter, Zhao H. Koh, Piengkwan Sribanditmongkol, Ben D. Fulcher, Jennifer M. Windt, Thomas Andrillon, and Naotsugu Tsuchiya. Beyond traditional sleep scoring: Massive feature extraction and data-driven clustering of sleep time series. *Sleep medicine*, 98:39–52, October 2022.
357. M. Diykh and Y. Li. Complex networks approach for EEG signal sleep stages classification. *Expert Systems with Applications*, 63:241–248, 2016.
358. R. Agarwal and J. Gotman. Computer-assisted sleep staging. *IEEE transactions on bio-medical engineering*, 48(12):1412–1423, December 2001.
359. J. Xu, H. Sun, Xing C, Fu X, Zhang Y, Zhang G, and Borjigin C. Sleep Analysis During Light Sleep Based on K-means Clustering and BiLSTM. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12999:207–214, 2021.
360. Wessam Al-Salman, Yan Li, Atheer Y. Oudah, and Sadiq Almaged. Sleep stage classification in EEG signals using the clustering approach based probability

distribution features coupled with classification algorithms. *Neuroscience research*, 188:51–67, March 2023.

361. M.U. Fadhlullah, A. Resahya, D.F. Nugraha, I.N. Yulita, Aisyah S, Megasari R, Kusumawaty D, Jupri A, Rusyati L, Rosjanuardi R, Hasanah L, Yulianti K, Wiji null, Samsudin A, and Nuraeni E. Sleep stages identification in patients with sleep disorder using k-means clustering. *Journal of Physics: Conference Series*, 1013(1), 2018.
362. Q. Pan, D. Brulin, and E. Campo. Wrist movement analysis for long-term home sleep monitoring. *Expert Systems with Applications*, 187, 2022.
363. Andreas M. Koupparis, Vasileios Kokkinos, and George K. Kostopoulos. Semi-automatic sleep EEG scoring based on the hypnospectrogram. *Journal of neuroscience methods*, 221:189–195, January 2014.
364. P. Van Hese, W. Philips, J. De Koninck, R. Van De Walle, and I. Lemahieu. Automatic detection of sleep stages using the EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 2: 1944–1947, 2001.
365. K.S. Prabhudesai, B.O. Mainsah, L.M. Collins, and C.S. Throckmorton. Augmented Latent Dirichlet Allocation (Lda) Topic Model with Gaussian Mixture Topics. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018:2451–2455, 2018.
366. A.M. Munk, K.V. Olesen, S.W. Gangstad, and L.K. Hansen. Semi-Supervised Sleep-Stage Scoring Based on Single Channel EEG. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018:2551–2555, 2018.
367. Mohammed Diykh, Yan Li, and Peng Wen. EEG Sleep Stages Classification Based on Time Domain Features and Structural Graph Similarity. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, 24(11):1159–1168, November 2016.
368. Sheng-Hsiou Hsu, Luca Pion-Tonachini, Jason Palmer, Makoto Miyakoshi, Scott Makeig, and Tzyy-Ping Jung. Modeling brain dynamic state changes with adaptive mixture independent component analysis. *NeuroImage*, 183:47–61, December 2018.
369. Muhammad Sohaib, Ayesha Ghaffar, Jungpil Shin, Md Junayed Hasan, and Muhammad Taseer Suleman. Automated Analysis of Sleep Study Parameters Using Signal Processing and Artificial Intelligence. *International journal of environmental research and public health*, 19 (20), October 2022.
370. Claus Metzner, Achim Schilling, Maximilian Traxdorf, Konstantin Tziridis, Andreas Maier, Holger Schulze, and Patrick Krauss. Classification at the accuracy limit: facing the problem of data ambiguity. *Scientific reports*, 12(1): 22121, December 2022.
371. M. Dutt, S. Redhu, M. Goodwin, and C.W. Omlin. Sleep Stage Identification based on Single-Channel EEG Signals using 1-D Convolutional Autoencoders. *2022 IEEE International Conference on E-Health Networking, Application and Services, HealthCom 2022*, pages 94–99, 2022.
372. H. Lerogeron, R. Picot-Clémente, L. Heutte, A. Rakotomamonjy, Jayne C, Mandic D, and Duro R. Learning an autoencoder to compress EEG signals via a neural network based approximation of DTW. *Procedia Computer Science*, 222:448–457, 2023.
373. Alexandra-Maria Tăuţan, Alessandro C. Rossi, Ruben de Francisco, and Bogdan Ionescu. Dimensionality reduction for EEG-based sleep stage detection: comparison of autoencoders, principal component analysis and factor analysis. *Biomedizinische Technik. Biomedical engineering*, 66(2):125–136, April 2021.
374. P. Moeynoi and Y. Kitjaidure. Hybrid dimensionality reduction of multi-sets using nature inspired algorithms and Discriminant Canonical Correlation Analysis for automatic sleep stage classification. *International Journal of Intelligent Engineering and Systems*, 12(1):277–289, 2019.
375. Jun Shi, Xiao Liu, Yan Li, Qi Zhang, Yingjie Li, and Shihui Ying. Multi-channel EEG-based sleep stage classification with joint collaborative representation and multiple kernel learning. *Journal of neuroscience methods*, 254:94–101, October 2015.
376. S. Sheykhivand, T. Yousefi Rezaii, A. Farzamnia, and M. Vazifehkhahi. Sleep stage scoring of single-channel EEG signal based on RUSBoost classifier. *Proceedings - 2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET 2018*, pages 36–42, 2018.
377. X. Mai and T. Yu. BootstrapNet: An Contrastive Learning Model for Sleep Stage Scoring based on Raw Single-Channel Electroencephalogram. *Proceedings - 2021 2nd International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2021*, pages 303–308, 2021.
378. Yiyang Zhang, Le Sun, Deepak Gupta, Xin Ning, and Prayag Tiwari. DCNet: A Self-supervised EEG Classification Framework for Improving Cognitive Computing-enabled Smart Healthcare. *IEEE journal of biomedical and health informatics*, January 2024.
379. S. Chang, Z. Yang, Y. You, and X. Guo. DSSNet: A Deep Sequential Sleep Network for Self-Supervised Representation Learning Based on Single-Channel EEG. *IEEE Signal Processing Letters*, 29:2143–2147, 2022.
380. Chaoqi Yang, Cao Xiao, M. Brandon Westover, and Jimeng Sun. Self-Supervised Electroencephalogram Representation Learning for Automatic Sleep Staging: Model Development and Evaluation Study. *JMIR AI*, 2 (1):e46769, January 2023.
381. L. Fraiwan and K. Lweesy. Neonatal sleep state identification using deep learning autoencoders. *Proceedings - 2017 IEEE 13th International Colloquium on Signal Processing and its Applications, CSPA 2017*, pages 228–231, 2017.
382. L. Duan, M. Li, C. Wang, Y. Qiao, Z. Wang, S. Sha, and M. Li. A Novel Sleep Staging Network Based on Data Adaptation and Multimodal Fusion. *Frontiers in Human Neuroscience*, 15, 2021.
383. Muhammad Zohaib Hassan Shah, Tengzi Liu, Yina Wei, and Dongping Yang. Unsupervised Feature Representation of Sleep EEG Data with Transient Deep Boltzmann Machine(). *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2023:1–5, July 2023.
384. Aaron Fisher, Brian Caffo, Brian Schwartz, and Vadim Zipunnikov. Fast, Exact Bootstrap Principal Component Analysis for $p > 1$ million. *Journal of the American Statistical Association*, 111(514):846–860, 2016.

385. Beth A. Lopour, Savas Tasoglu, Heidi E. Kirsch, James W. Sleigh, and Andrew J. Szeri. A continuous mapping of sleep states through association of EEG with a mesoscale cortical model. *Journal of computational neuroscience*, 30(2):471–487, April 2011.
386. V. Kumar, L. Reddy, S. Kumar Sharma, K. Dadi, C. Yarra, R.S. Bapi, S. Rajendran, Wang L, Dou Q, Fletcher P.T, Speidel S, and Li S. mulEEG: A Multi-view Representation Learning on EEG Signals. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13433:398–407, 2022.
387. B.A. Savareh, A. Bashiri, A. Behmanesh, G.H. Meftahi, and B. Hatef. Performance comparison of machine learning techniques in sleep scoring based on wavelet features and neighboring component analysis. *PeerJ*, 2018(7), 2018.
388. R.N. Khushaba, R. Elliott, A. AlSukker, A. Al-Ani, and S. McKinley. Orthogonal locality sensitive fuzzy discriminant analysis in sleep-stage scoring. *Proceedings - International Conference on Pattern Recognition*, pages 165–168, 2010.
389. H.X. Mao, J. Widjaja, Y. Guo, J. Yin, and R. Vinjamuri. Finding Robust Low Dimensional Features for Sleep Detection Using EEG Data. *2022 IEEE 2nd International Conference on Data Science and Computer Application, ICDSCA 2022*, pages 42–45, 2022.
390. Cabir Vural and Murat Yildiz. Determination of sleep stage separation ability of features extracted from EEG signals using principle component analysis. *Journal of medical systems*, 34(1):83–89, February 2010.
391. Claus Metzner, Achim Schilling, Maximilian Traxdorf, Holger Schulze, Konstantin Tziridis, and Patrick Krauss. Extracting continuous sleep depth from EEG data without machine learning. *Neurobiology of sleep and circadian rhythms*, 14:100097, May 2023.
392. Shirin Enshaeifar, Samaneh Kouchaki, Clive Cheong Took, and Saeid Sanei. Quaternion Singular Spectrum Analysis of Electroencephalogram With Application in Sleep Analysis. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, 24 (1):57–67, January 2016.
393. E.M. Petrova, V.G. Rybin, T.I. Karimov, Shaposhnikov S, and Saint Petersburg Saint Petersburg Electrotechnical University "LETI", Prof. Popov Str. 5. Deep Transfer Learning for Sleep Stages Classification by EEG Data. *Proceedings of the Seminar on Digital Medical and Environmental Systems and Tools, DMEST 2023*, pages 106–108, 2023.
394. J. Ye, Q. Xiao, J. Wang, H. Zhang, J. Deng, and Y. Lin. <italic>CoSleep</italic>: A Multi-View Representation Learning Framework for Self-Supervised Learning of Sleep Stage Classification. *IEEE Signal Processing Letters*, 29: 189–193, 2022.
395. A. Salazar, L. Vergara, and R. Miralles. On including sequential dependence in ICA mixture models. *Signal Processing*, 90(7):2314–2318, 2010.
396. A.K. Dutta, Y. Albagory, M. Al Faraj, Y.A.M. Eltahir, and A.R.W. Sait. Optimal Sparse Autoencoder Based Sleep Stage Classification Using Biomedical Signals. *Computer Systems Science and Engineering*, 44(2):1517–1529, 2023.
397. J. Zhang, Y. Wu, J. Bai, and F. Chen. Automatic sleep stage classification based on sparse deep belief net and combination of multiple classifiers. *Transactions of the Institute of Measurement and Control*, 38(4):435–451, 2016.
398. M. Vaezi and M. Nasri. AS3-SAE: Automatic Sleep Stages Scoring Using Stacked Autoencoders. *Frontiers in Biomedical Technologies*, 10(4):400–416, 2023.
399. R.K. Tripathy and U. Rajendra Acharya. Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework. *Biocybernetics and Biomedical Engineering*, 38(4):890–902, 2018.
400. Orestis Tsinalis, Paul M. Matthews, and Yike Guo. Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders. *Annals of biomedical engineering*, 44(5):1587–1597, May 2016.
401. S. Najdi, A.A. Gharbali, J.M. Fonseca, Monte da Caparica Caparica FCT-Department of Electrical Engineering, Universidade Nova de Lisboa, Camarinha-Matos L.M, Parreira-Rocha M, and Ramezani J. Feature transformation based on Stacked Sparse Autoencoders for sleep stage classification. *IFIP Advances in Information and Communication Technology*, 499: 191–200, 2017.
402. Y. Wang, Y. Wang, L. Yao, and X. Zhao. Single Channel Sleep Staging Based on Unsupervised Feature Learning. *10th International Conference on Intelligent Control and Information Processing, ICICIP 2019*, pages 180–183, 2019.
403. Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of neural engineering*, 18(4), March 2021.
404. Z. Chen, N. Ono, M.D. Altaf-Ul-Amin, S. Kanaya, M. Huang, Park T, Cho Y.-R, Hu X.T, Yoo I, Woo H.G, Wang J, Facelli J, Nam S, and Kang M. IVAE: An Improved Deep Learning Structure for EEG Signal Characterization and Reconstruction. *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, pages 1909–1913, 2020.
405. G.-R. Liu, Y.-L. Lo, Y.-C. Sheu, and H.-T. Wu. Explore Intrinsic Geometry of Sleep Dynamics and Predict Sleep Stage by Unsupervised Learning Techniques. *Springer Optimization and Its Applications*, 168:279–324, 2021.
406. Kirubin Pillay, Anneleen Dereymaeker, Katrien Jansen, Gunnar Naulaers, Sabine Van Huffel, and Maarten De Vos. Automated EEG sleep staging in the term-age baby using a generative modelling approach. *Journal of neural engineering*, 15(3):036004, June 2018.
407. Farid Yaghouby, Pradeep Modur, and Sridhar Sunderam. Naive scoring of human sleep based on a hidden Markov model of the electroencephalogram. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2014: 5028–5031, 2014.
408. Hojat Ghimatgar, Kamran Kazemi, Mohammad Sadegh Helfroush, Kirubin Pillay, Anneleen Dereymaker, Katrien Jansen, Maarten De Vos, and Ardalan Aarabi. Neonatal EEG sleep stage classification based on deep learning and HMM. *Journal of neural engineering*, 17(3):036031, June 2020.
409. A.B. Rossow, E.O.T. Salles, and K.F. Côco. Automatic sleep staging using a single-channel EEG modeling by

Kalman Filter and HMM. *2011 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living, BRC 2011*, pages 134–139, 2011.

410. S. Riazy, T. Wendler, J. Pilz, M. Glos, T. Penzel, Maciel C.D, Fred A, Gamboa H, and Vaz M. Heuristic approximation of the map estimator for automatic two-channel sleep staging. *BIOSIGNALS 2017 - 10th International Conference on Bio-Inspired Systems and Signal Processing, Proceedings; Part of 10th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2017*, 4:236–241, 2017.
411. C. Ren, L. Sun, and D. Peng. A Contrastive Predictive Coding-Based Classification Framework for Healthcare Sensor Data. *Journal of Healthcare Engineering*, 2022, 2022.
412. H. Lee, E. Seong, D.-K. Chae, De Raedt L, and De Raedt L. Self-Supervised Learning with Attention-based Latent Signal Augmentation for Sleep Staging with Limited Labeled Data. *IJCAI International Joint Conference on Artificial Intelligence*, pages 3868–3876, 2022.
413. Jingcong Li, Fei Wang, Haiyun Huang, Feifei Qi, and Jiahui Pan. A novel semi-supervised meta learning method for subject-transfer brain-computer interface. *Neural networks : the official journal of the International Neural Network Society*, 163:195–204, June 2023.
414. H. Banville, G. Moffat, I. Albuquerque, D.-A. Engemann, A. Hyvarinen, and A. Gramfort. Self-Supervised Representation Learning from Electroencephalography Signals. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2019, 2019.
415. Y. Zou, Y. Zhang, and X. Zhao. Self-Supervised Time Series Classification Based on LSTM and Contrastive Transformer. *Wuhan University Journal of Natural Sciences*, 27(6):521–530, 2022.

Appendix 1

Reference	Method	Role	Data Type	Data Set	Population	#	Duration	Metric	Value
Zhang et al., 2023 [156]	Adaptive Memory Network With Self-supervised Learning	Classification	EEG	CAP database	OSA patients and control group	108	1 night	Accuracy	0.98
Chao et al., 2006 [328]	Adaptive Fuzzy Clustering	Classification	EEG	Sleep-EDF	Adults with mild sleep difficulty	1	1 night	-	-
Hermawan et al., 2013 [122]	Adaptive Multicodebook Fuzzy Neuro Generalized Learning with Decision Tree-based Clustering	Pre-classificiation Clustering	ECG	MIT-BIH and self-collected	Healthy adults	21	1 night	Accuracy	0.81 - 0.82
Dursun et al., 2012 [329]	Artificial Immune Clustering	Train Set Clustering	EEG	Self-collected	Healthy adults	6	1 night	Accuracy	0.81
Prabhudesai, Collins and Mainsah, 2019 [330]	Autoencoder and LDA Topic Model	Classification	EEG	Sleep-EDF	Healthy adults	30	1 night	Fowlkes-Mallows score	0.71
Wang et al., 2014 [331]	Collective Dynamical Modeling Clustering	Classification	EEG	Day Kimball Hospital dataset	Suspected sleep disorders	200	1 night	Accuracy	0.94
Zhang et al., 2018 [332]	Complex-Valued Unsupervised Convolutional Neural Network	Classification	EEG	UCD and MIT-BIH	SDB patients	41	1 night	Accuracy	0.87
Yu et al., 2019 [333]	Density-Distance-based K-Means	Classification	EEG	Sleep-EDF	Healthy adults	6	1 night	Accuracy	0.74
Zheng et al., 2020 [334]	Empirical Mode Decomposition and K-Means	Classification	EEG	Sleep-EDF	Healthy adults	3	1 night	Accuracy	0.83
Mporas et al., 2015 [185]	Expectation Maximization Algorithm	Train Set Clustering	EEG	Sleep-EDF	Healthy adults	61	1 night	Accuracy	0.93
Obayya and Abou-Chadi, 2014 [335]	Fuzzy C-Means Clustering	Classification	EEG	Cairo Center for Sleep Disorders	Healthy adults	24	1 night	Accuracy	0.92
Raiesdana et al., 2013 [336]	Fuzzy C-Means Clustering	Classification	EEG	Self-Collected	Healthy adults	5	1 night	Accuracy	0.81
Yulita et al., 2017 [337]	Fuzzy C-Means Clustering	Feature Extraction	EEG	UCD	OSA patients	25	1 night	F1 Score	0.73
Kumar, 1977 [51]	Fuzzy C-Means Clustering	Classification	EEG	Self-collected	-	12	3-4 nights	Accuracy	0.9
Gath et al., 1989 [53]	Fuzzy C-Means Clustering	Classification	EEG	Self-collected	Healthy adults	10	1 night	-	-
Gath et al., 1994 [56]	Fuzzy C-Means Clustering	Classification	EEG	Self-collected	Healthy and sleep disordered adults	8	1 night	-	-
Chao et al., 2006 [339]	Fuzzy C-Means Clustering	Classification	EEG	Sleep-EDF	Healthy adults	1	1 night	-	-
Yin et al., 2021 [184]	Fuzzy Few Samples Clustering	Train Set Clustering	EEG	-	-	-	-	Accuracy	0.85
Nasibov et al., 2010 [340]	Fuzzy Neighborhood and Density-based Spatial Clustering	Classification	EEG	-	-	1	25 min	-	-
Gath et al., 1980 [54]	Fuzzy Subset Theory and Optimal Fuzzy Partition	Classification	EEG, EOG and EMG	Self-collected	Healthy adults	-	1 night	-	-
Rajalakshmi et al., 2018 [341]	Gaussian Mixture Model	Classification	EEG	Sleep-EDF	-	50	3 hours	Accuracy	0.89
Koivuluoma et al., 2000 [342]	Gaussian Mixture Model	Classification	EOG and EEG	SIESTA database	Healthy adults	7	1 night	Accuracy	0.72 - 0.90
Garcia-Molina et al., 2013 [343]	Gaussian Mixture Model	Classification	EEG	Self-collected	healthy adults	10	1 night	Cohen's Kappa	0.79
Acharya et al., 2010 [344]	Gaussian Mixture Model	Classification	EEG	UCD	Suspected sleep disorders	37	1 night	Accuracy	0.89
Rajalakshmi et al., 2018 [345]	Gaussian Mixture Model	Classification	EEG	Sleep-EDFx	Healthy adults	61	1 night	Accuracy	0.87
Guo et al., 2023 [346]	Gaussian Mixture Model and K-Means	Pre-classificiation Clustering	EEG	Sleep-EDFx	Healthy adults	20	1 night	Accuracy	0.84
Lajnef et al., 2015 [347]	Hierarchical Clustering	Ensemble Building	EEG, EMG and EOG	Self-collected	Healthy adults	15	1 night	Accuracy	0.88
Gerla et al., 2009 [348]	Hierarchical Clustering	Classification	EEG, EOG, EMG, ECG and Respiration	Self-collected	Newborns	10	1 night	-	-
Escola et al., 1991 [58]	Hierarchical Clustering	Classification	EEG	Self-Collected	Healthy adults	-	1 night	Accuracy	0.80
Gerla et al., 2019 [349]	Hierarchical Clustering with Ward's Algorithm	Classification	EEG, EOG and EMG	Self-collected	Insomniacs and healthy adults	36	1 night	F1 Score	0.88
Gerla et al., 2018 [350]	Hybrid Hierarchical Clustering	Classification	EEG	Self-collected	-	2	1 night	-	-
Larsen et al., 1980 [60]	Isodata Clustering	Classification	EEG	Self-collected	Healthy adults	-	4 hours	-	-
Kazemi et al., 2022 [351]	Iterative Binary Clustering	Classification	EEG	Sleep-EDF and CAP database	Healthy adults and RBD patients	40	1 night	Accuracy	0.73
Rodríguez-Sotelo et al., 2014 [352]	J Means Clustering	Classification	EEG	Physionet	Healthy adults	39	1 night	Accuracy	0.80
Rodríguez-Sotelo et al., 2017 [353]	J-Means+ Clustering	Classification	EEG	Self-collected	Healthy adults	20	1 night	Accuracy	0.73
Shao et al., 2017 [354]	K-Means Clustering	Classification	EEG	CAP database	-	3	1 night	Accuracy	0.64 - 0.76
Wang et al., 2019 [355]	K-Means Clustering	Clustering	EEG	Physionet	Healthy adults	39	1 night	Accuracy	0.82
Decat et al., 2022 [356]	K-Means Clustering	Classification	EEG	Cleveland Children's Sleep and Health Study	Teenagers	12	1 night	-	-
Diykh et al., 2016 [357]	K-Means Clustering	Classification	EEG	UCD	Healthy and sleep disordered adults	25	2 night	Accuracy	0.92
Agarwal et al., 2001 [358]	K-Means Clustering	Classification	EEG	Self-Collected	Healthy and sleep disordered adults	12	1 night	Accuracy	0.80
Krajca et al., 2018 [187]	K-Means Clustering	Classification	ECG	Self-collected	Newborns	20	20 min -2 hours	-	-

Table 10. Publications on Sleep Staging Pt.1

Reference	Method	Role	Data Type	Data Set	Population	#	Duration	Metric	Value
Güneş et al., 2010	K-Means Clustering	Weighting Features	EEG	Self-collected	Healthy adults	5	1 night	Accuracy	0.82
Xu et al., 2021	K-Means Clustering	Feature Extraction	EEG	Sleep-EDFx	-	-	1 night	Accuracy	0.81
Al-Salman et al., 2023	K-Means Clustering	Feature Extraction	EEG	Sleep-EDF	Healthy adults	8	1 night	Accuracy	0.97
Fadhlullah et al., 2018	K-Means Clustering	Classification	EEG	UCD	Suspected SDB	25	1 night	-	-
Pan et al., 2022	K-Means Clustering	Classification	Accelerometer	Self-collected	Healthy adults	5	1 night	Accuracy	0.75 - 0.97
Koupparis et al., 2014	K-Means Clustering	Classification	EEG Spectrograms	Self-collected	Healthy adults	10	1 night	Cohen's Kappa	0.61
Van Hese et al., 2001	K-Means Clustering	Classification	EEG	Self-collected	-	1	1 night	-	-
Prabhudesai et al., 2018	Latent Dirichlet Allocation and Gaussian Mixture Model	Classification	EEG	Sleep-EDF	Healthy adults	30	1 night	Fowlkes-Mallows score	0.75
Pan et al., 2021	MK-Means Clustering	Classification	Accelerometer	Self-collected	Healthy adults	1	1 night	-	-
Gath et al., 1989	Optimal Fuzzy Clustering	Classification	EEG	Self-collected	Healthy adults	1	1 night	-	-
Katayama et al., 1995	Self-Organizing Map	Classification	EEG	Self-collected	Healthy adults	1	1 night	Match Ratio	0.79
Munk et al., 2018	Semi-Supervised Gaussian Mixture Model	Classification	EEG	Sleep-EDFx	Healthy adults	19	2 nights	Accuracy	0.7
Diykh et al., 2016	Structural Graph Similarity and K-Means	Classification	EEG	Sleep-EDF and MONS-TCTS	Healthy adults	16	1 night	Accuracy	0.96
Tian et al., 2017	Proportion-based Clustering	Classification	EEG	Sleep-EDF	Healthy adults and adults with mild difficulty falling asleep	20	20 hours	Accuracy	0.91
Gath et al., 1983	Time-Dependent Fuzzy Clustering	Classification	EEG	Self-collected	Healthy adults	5	2 Or 3 nights	-	-
Hsu et al., 2018	Adaptive Mixture Independent Component Analysis	Separate Components	EEG	CAP database	Healthy adults	17	1 night	Accuracy	0.75
Sohaib et al., 2022	Autoencoder	Classification	EEG	Sleep-EDF	Healthy adults	153	1 night	Accuracy	0.99
Metzner et al., 2022	Autoencoder	Feature Extraction	EEG	Self-collected	Healthy adults	68	1 night	-	-
Dutt et al., 2022	Autoencoder	Reconstruction	EEG	Sleep-EDF	Healthy adults	20	2 nights	Accuracy	0.87
Lerogeron et al., 2023	Autoencoder	Feature Extraction	EEG	Sleep-EDF, SHHS and DOD-O	Various cohorts with and without sleep disorders	462	1 night	Accuracy	0.53 - 0.67
Kim et al., 2024	Autoencoder	Pretraining	EEG	Sleep-EDF	Healthy adults	20	1 night	Accuracy	0.84
Tăuţan et al., 2021	Autoencoder, Principal Component Analysis and Factor Analysis	Feature Extraction	EEG, EMG, ECG and Respiratory Signals	PhysioNet	-	994	1 night	Accuracy	0.95
Moeynoi et al., 2019	Canonical Correlation Analysis	Feature Extraction	EEG	Self-collected	Healty adults and adults with various pathologies	-	-	Accuracy	0.98
Shi et al., 2015	Collaborative Representation	Classification	EEG	UCD	Suspected SDB	25	1 night	Accuracy	0.80
Sheykhivand et al., 2018	Complete Ensemble Empirical Mode Decomposition	Feature Extraction	EEG	Sleep-EDF	Healthy adults and sleepy adults	8	1 night	Accuracy	0.91
Mai et al., 2021	Contrastive Learning	Classification	EEG	Sleep-EDF	Healthy adults	20	1 night	Accuracy	0.86
Zhang et al., 2024	Contrastive Learning	Classification	EEG	Sleep-EDF	-	-	1 night	Accuracy	0.81
Chang et al., 2022	Contrastive Learning	Classification	EEG	Sleep-EDF and ISRUC	Healthy adults	120	1 night	Accuracy	0.80
Yang et al., 2023	Contrastive Learning	Feature Extraction	EEG	Sleep-EDF, SHHS and MGH dataset	Healthy adults	10882	Varying Numbers Of nights	Accuracy	0.72 - 0.87
Han et al., 2022	Cross-Modal Contrastive Hashing Retrieval	Classification	Infrared Video	Self-collected	Suspected Sleep Disorders	105	1 night	Accuracy	0.81
Fraiwan et al., 2017	Deep Autoencoder	Classification	EEG	EEGdat dataset	Newborns	29	2-3 hours	Accuracy	0.80
Yulita et al., 2017	Deep Belief Network	Feature Extraction	EEG	UCD	Sleep disordered adults	25	1 night	F1 score	0.75
Reimer et al., 2018	Deep Belief Network	Feature Extraction	Chest Strap and Accelerometer	Self-collected	Healthy adults	19	1 night	Accuracy	0.83
Duan et al., 2021	Deep Belief Network with Restricted Boltzmann Machines	Classification	EEG	Sleep-EDF and self-collected	-	217	1 night	Accuracy	0.84 - 0.88
Hassan Shah MZ et al., 2023	Deep Boltzmann Machine	Classification	EEG	Sleep-EDF	-	-	1 night	F1 Score	0.96
Mendonça et al., 2019	Deep Stacked Autoencoder	Classification	ECG	CAP database, UCD and self-collected	Healthy adults and SDB patients	158	1 night	Accuracy	0.77
Fisher et al., 2016	Fast Bootstrap Principal Component Analysis	Feature Extraction	EEG	SHHS	Healthy adults	392	1 night	-	-
Falie et al., 2010	Independent Component Analysis	Signal Processing	3D camera	Self-Collected	Healthy adults	1	1 night	-	-
Chen et al., 2021	Latent Dirichlet Allocation	Clustering	EEG	SHHS	healthy adults	5736	1 night	-	-
Lopour et al., 2011	Local Linear Embedding	Classification	EEG	Sleep-EDF	Healthy adults	4	1 night	-	-
Kumar et al., 2022	Multi-View Self-Supervised Method	Classification	EEG	Sleep-EDF, SHHS	Healthy adults	404	1 night	Accuracy	0.78 - 0.81
Savareh et al., 2018	Neighboring Component Analysis	Feature Extraction	EEG, EOG	Sleep-EDF	-	61	1 night	Accuracy	0.90
Khushaba et al., 2010	Orthogonal Locality Sensitive Fuzzy Discriminant Analysis	Feature Extraction	EEG, EMG and EOG	Self-collected	ICU patients	9	1 night	Accuracy	0.91
Mao et al., 2022	Principal Component Analysis and Autoencoder	Feature Extraction	EEG	Physionet	Healthy adults	1	1 night	Accuracy	0.90
Vural et al., 2010	Principal Component Analysis	Feature Extraction	EEG	PhysioNet	-	-	-	Accuracy	0.34 - 0.97
Metzner et al., 2023	Principal Component Analysis	Feature Extraction	EEG	Self-collected	Healthy adults	68	1 night	General Discrimination Values	−0.21

Table 11. Publications on Sleep Staging Pt.2

Reference	Method	Role	Data Type	Data Set	Population	#	Duration	Metric	Value
Enshaeifar et al., 2016	Quaternion-Based Singular Spectrum Analysis	Classification	EEG	Self-collected	Healthy adults	36	1 Week	Sensitivity	0.68
Petrova et al., 2023	Recurrent Autoencoder	Classification	EEG	Sleep-EDF and UCR Archive	Healthy adults	150	1 night	Accuracy	0.85
Ye et al., 2022	Self-supervised Representation Learning	Classification	EEG	Sleep-EDF and ISRUC	Healthy adults and SDB adults	139	1 night	Accuracy	0.58 - 0.71
Ramnath and Katkoori, 2020	Self-supervised Temporal Feature Learning with Autoencoder and Gaussian Mixture Model	Classification	Accelerometer and PPG	Self-collected	Healthy adults	39	7-14 nights	Accuracy	0.63
Salazar et al., 2010	Sequential Independent Analysis Mixture Models	Classification	EEG	-	OSA patients	2	2 night	Error rate	0.09 - 0.2
Dutta et al., 2023	Sparse Autoencoder With Smoothed Regularization	Classification	EEG	Sleep-EDF	Healthy adults	-	1 night	Accuracy	0.99
Zhang et al., 2016	Sparse Deep Belief Net	Classification	EEG	UCD	SDB patients	25	1 night	Accuracy	0.91
Vaezi and Nasri, 2023	Stacked Autoencoder	Classification	EEG and ECG	SHHS and ISRUC	Healthy adults	40	1 night	Accuracy	0.94 - 0.96
Wei et al., 2018	Stacked Autoencoder	Pretraining	EEG	MIT-BIH	Healthy adults	18	1 night	Accuracy	0.77
Tripathy et al., 2018	Stacked Autoencoder	Classification	EEG and ECG	MIT-BIH	-	18	1 night	Accuracy	0.86 - 0.96
Tsinalis et al., 2016	Stacked Sparse Autoencoder	Classification	EEG	Physionet	Healthy adults	20	1 night	Accuracy	0.78
Najdi et al., 2017	Stacked Sparse Autoencoder	Classification	EEG	ISRUC-Sleep	Healthy adults	9	1 night	Accuracy	0.82
Wang et al., 2019	Symmetric Convolutional Neural Network	Classification	EEG	Sleep-EDF	Healthy Adults	6	1 night	Accuracy	0.77
Banville et al., 2021	Temporal Context Prediction and Contrastive Predictive Coding	Classification	EEG	Physionet Challenge 2018 dataset and TUH Abnormal EEG dataset	OSA patients and healthy adults	3323	1 night and Short Time Measurement	Accuracy	0.83
Chen et al., 2020	Variational Autoencoder	Feature Extraction	EEG	UCD and MIT-BIH	Healthy adults	41	1 night	Accuracy	0.68
Takeda et al., 2015	Gibbs Sampling	Classification	HRV	SHRSV database	Healthy adults	45	1 night	Accuracy	0.76
Liu et al., 2021	Diffusion Map, Alternating Diffusion and Hidden Markov Model	Classification	EEG	Sleep-EDFx	Healthy adults and slightly sleep disordered adults	42	1 night	Accuracy	0.77 - 0.83
Pillay et al., 2018	Hidden Markov Model and Gaussian Mixture Model	Classification	EEG	NICU of the University Hospitals, Leuven, Belgium	Newborns	16	1 night	Accuracy	0.86
Yaghouby et al., 2014	Hidden Markov Model	Classification	EEG	Physionet	Healthy adults	22	1 night	Cohen's Kappa	0.7
Ghimatgar et al., 2020	Hidden Markov Model and Modified Graph Clustering Ant Colony Optimization)	Post processing	EEG	NICU of the University Hospitals, Leuven, Belgium	Newborns	16	1 night	Accuracy	0.79 - 0.82
Trevenen et al., 2019	Hidden Markov Models snd Generalized Linear Mixed Models	Classification	Smartwatch Accelerometer	Western Australian Pregnancy Cohort Study	Healthy adults	242	1 night	F1 Score	0.432
Rossow et al., 2011	K-Means Segmental Hmm	Classification	EEG	MIT-BIH	-	-	-	Accuracy	0.60
Riazy et al., 2017	Maximum A Posteriori Estimation	Classification	EEG	Self-collected	Healthy adults	5	1 night	Error rate	0.25
Vanbuis et al., 2022	Viterbi Hidden Markov Model	Classification	Tracheal sound sensor, Heartrate, RIP belts, Flow and SpO2	Self-collected	Patients with SDB	400	1 night	Accuracy	0.79
Heremans et al., 2022	Adversarial Domain Adaptation	Domain Adaption	EEG	Surrey-cEEGrid dataset, Dreem-Headband dataset, Leuven-CBTE-46 and MASS	Various cohorts with and without sleep disorders	282	1 night	Accuracy Improvement	0.07 - 0.27
Zhang et al., 2021	Competition Convolutional Neural Network	Classification	EEG	UCD and Sleep-EDF	SDB patients	67	1 night	Accuracy	0.77 - 0.83
Zhao et al., 2021	Conditional and Collaborative Adversarial Domain Adaptation	Domain Adaption	EEG	Sleep-EDF	Healthy adults	96	1 night	Accuracy	0.85
Xiao et al., 2021	Contrastive Learning	Classification	EEG	Sleep-EDF	Healthy adults	20	1 night	Accuracy	0.701
Ren et al., 2022	Contrastive Predictive Coding	Classification	EEG, EOG and EMG	Physiobank	Healthy adults	197	1 night	Accuracy	0.973
Gao et al., 2023	Domain Adversarial Neural Networks and Domain Self-Attention	Domain Adaption	EEG	Sleep-EDFx and self-collected	Healthy adults and sleep disordered adults	41	1-2 nights	Accuracy	0.80
Fan et al., 2022	Domain Statistics Alignment	Domain Adaption	EEG	MASS, SHHS, Sleep-EDF, UCD and HSFU	Various cohorts with and without sleep disorders	6399	1 night	Accuracy	0.81
Amor et al., 2022	Hierarchical Multi-Agent System	Classification	EEG	Physionet, MASS and Sahloul hospital	Healty adults and adults with various pathologies	4	1 night	Accuracy	0.94
Lee et al., 2022	Self-Supervised Learning with Attention-Aided Positive Pairs	Feature Extraction	EEG	Sleep-EDFx and ISRUC	Healthy adults	163	1 night	Accuracy	0.89
Li et al., 2023	Semi-Supervised Meta Learning	Classification	EEG	Sleep-EDF	Healthy adults	20	1 night	Accuracy	0.83
Banville et al., 2019	Temoral Contrastive Tasks	Feature Extraction	EEG	Sleep-EDF andMASS	Healthy adults	145	1-2 nights	Accuracy	0.77
Luo et al., 2023	Time-Series Domain Adaptation	Domain Adaption	EEG	SLEEP-EDF, SHHS, and ISRUC-Sleep	Various cohorts with and without sleep disorders	311	1 night	Improved Accuracy	0.52 - 0.14
He et al., 2023	Unsupervised Domain Adaptation	Domain Adaption	EEG	Sleep-EDF, SHHS and Physionet		220	1-2 nights	Accuracy Improvement	0.03
Yoo et al., 2022	Unsupervised Domain Adaptation	Domain Adaption	EEG	MASS, Sleep-EDF and Sleep-EDF-st	Various cohorts with and without sleep disorders	242	1 night	Accuracy	0.49 - 0.93
Eldele et al., 2023	Unsupervised Domain Adaption	Domain Adaption	EEG	Sleep-EDF, SHHS-1 and SHHS-22	Healthy and sleep disordered adults	106	1 night	Accuracy	0.74
Zou et al., 2022	Unsupervised Long Short-Term Memory and Contrastive Transformer-based Model	Classification	EEG	Sleep-EDF	-	-	-	Accuracy	0.84

Table 12. Publications on Sleep Staging Pt.3

Appendix B

Publication II

Data Mining and Knowledge Discovery (2024) 38:976–1005
https://doi.org/10.1007/s10618-023-00985-x

Anomaly detection in sleep: detecting mouth breathing in children

Luka Biedebach[1] · María Óskarsdóttir[1] · Erna Sif Arnardóttir[1] · Sigridur Sigurdardóttir[1] · Michael Valur Clausen[2] · Sigurveig Þ. Sigurdardóttir[2] · Marta Serwatko[2] · Anna Sigridur Islind[1]

Received: 17 May 2022 / Accepted: 3 October 2023 / Published online: 13 November 2023

Abstract

Identifying mouth breathing during sleep in a reliable, non-invasive way is challenging and currently not included in sleep studies. However, it has a high clinical relevance in pediatrics, as it can negatively impact the physical and mental health of children. Since mouth breathing is an anomalous condition in the general population with only 2% prevalence in our data set, we are facing an anomaly detection problem. This type of human medical data is commonly approached with deep learning methods. However, applying multiple supervised and unsupervised machine learning methods to this anomaly detection problem showed that classic machine learning methods should also be taken into account. This paper compared deep learning and classic machine learning methods on respiratory data during sleep using a leave-one-out cross validation. This way we observed the uncertainty of the models and their performance across participants with varying signal quality and prevalence of mouth breathing. The main contribution is identifying the model with the highest clinical relevance to facilitate the diagnosis of chronic mouth breathing, which may allow more affected children to receive appropriate treatment.

Keywords Anomaly detection · Sleep · Machine learning · Mouth breathing

1 Introduction

The rise of machine learning in sleep research is already revolutionizing the diagnosis of sleep disorders (Arnardottir et al 2021) with the automatic classification of sleep stages (Korkalainen et al 2019) and the detection of respiratory events (Huang and Ma 2021). Machine learning as a part of sleep research and clinical practice can reduce the manual effort of physicians and sleep technologists and

Responsible editor: Johannes Fürnkranz.

Extended author information available on the last page of the article

increase the well-being of the patient through more precise diagnosis and less invasive sleep measurements (Biedebach et al 2023). Acquiring labels for one night of sleep recording requires 2–3 h of manual review by a sleep technologist who is an expert in the field, which is both time-consuming and expensive (Arnardottir et al 2022). Some potentially important events during sleep, such as mouth breathing, are often not labeled at all. Mouth breathing is a particular problem for children, since they may face developmental issues if they breath through their mouth during sleep, at an early age (Gozal 1998). Children with sleep-disordered breathing can suffer from serious long-term implications if their condition is not recognized and appropriately treated (Marcus 2001). In fact, chronic mouth breathing can lead to obstructive sleep apnea (Izu et al 2010), learning disorders, (Fensterseifer et al 2013) and a malformation of the child's jaw area (Denotti et al 2014). The task of identifying mouth breathing with machine learning is challenging, since we are facing an anomaly detection problem. Healthy children usually breathe through their nose, which makes mouth breathing an anomalous behavior (Lee et al 2015). In this paper, we analyze mouth breathing of children, in a highly imbalanced data set, which included only few mouth breathing sequences because most of the children did not breathe through the mouth at all. From the 20 labeled recordings, the child with the highest duration of mouth breathing had a total length of 1980 s or 33 min of mouth breathing in a night with 10 h of sleep. Overall, the data set that has only 2.4% positive examples. Therefore, we assume that mouth breathing is an anomaly. In this paper, we aimed to reduce the manual effort for identifying mouth breathing in order to enable a more efficient diagnosis of the condition, which will hopefully enable more children to receive treatment before these health implications surface. There are two central challenges to identifying mouth breathing. Firstly, even for the sleep technologist, it is difficult to make a clear distinction between mouth and nose breathing, since the boundary is blurred. In general, there is no mouth breathing, if the mouth is closed and the air is solely flowing through the nose. However, the same generalization is not valid for the other way around; air can flow both through the nose and the mouth when there is mouth breathing (Koutsourelakis et al 2006). Secondly, it is challenging to acquire a sufficient amount of labeled mouth breathing events, since mouth breathing is an anomalous behavior in the healthy population and they are usually not labeled. As a result, the labeled recording might include only a few or no mouth breathing events at all.

When approaching this type of human medical data with machine learning, different rules than for tabular data apply. We need to consider that each training example is a breathing sequence that belongs to a certain unique individual. This impacts both the training and testing of the machine learning model. Splitting the data with a common random train test split, could lead to a data leakage problem during the model training. Peralta et al (2021) show in a systematic review in machine learning for deep brain stimulation, that more than half of the papers in this field do not do a patient-wise validation. Therefore, we created the train, test and validation set by separating the data by participants as shown by Oner et al (2020). This is in line with the practical aspects of implementing the machine learning model in clinical practice, where the data of a new patient is fully separated from the data the model was trained on. The same logic holds for the evaluation of a machine learning model

with human medical data. Evaluating the performance of the model on participants separately can reveal whether the model performance varies among different groups of participants with certain characteristics and whether the model can generalize on all patients.

In this paper, we aimed to find the best way to predict pediatric mouth breathing during sleep by comparing different supervised and unsupervised machine learning models. The data set included sleep recordings of 111 participants using oronasal cannulas. We transformed the data set, by first splitting the full sleep recordings into 10-second subsequences. We trained the model on multiple signals of the sleep recordings including thorax and abdomen movement, oral pressure, nasal pressure, blood oxygen saturation, heart rate, audio volume and position. We tested the model using a leave-one-out cross validation and chose the model with the highest clinical relevance. The main contributions of this paper are three-fold: (i) to illustrate the challenges and required preprocessing steps for applying machine learning to sleep data, (ii) to identify the models with highest clinical value, and (iii) to contribute to the discourse of when deep learning is needed and when simplicity is key.

Our work makes an important contribution to the field of sleep research, as we show that mouth breathing during sleep can be automatically identified with machine learning, which allows a faster diagnosis of mouth breathing. Most significantly, our work contributes to machine learning as we show the challenges of working with human medical data and outline a sensible preprocessing, training and evaluation method to counteract them. Importantly, we approached this problem with different machine learning methods, including a naive baseline, a classic machine learning model, time series classifiers, deep learning models and an unsupervised anomaly detection method. Evaluating these different methods in a leave-one-out cross validation showed their performance across the whole population and on an individual basis, which raised the question whether classic methods are preferable over deep learning for anomaly detection in sleep. The rest of this paper is organized as follows.

In the next section, we summarize the existing literature related to unsupervised anomaly detection and mouth breathing identification. Then, we describe our proposed methodology for automatic detection of mouth breathing events, followed by a presentation of our results. The paper ends with a discussion of the implications of our contribution and steps for future work.

2 Related work

2.1 Time series anomaly detection

Time series anomaly detection, as a subfield of anomaly detection, has been studied in literature. Common application fields of anomaly detection are healthcare (Chauhan and Vig 2015), financial fraud (Fu et al 2016), robotics (Park et al 2018) or network intrusion (Leung and Leckie 2005). Literature reviews (Blázquez-García et al 2021; Chalapathy and Chawla 2019) show the broad range of methods and application fields of time series anomaly detection. Experimental comparisons have been

conducted and compare the performance of both supervised (Freeman et al 2021) and unsupervised (Rewicki et al 2023) anomaly detection methods on time series bench marking data sets. This paper did not aim to do a exhaustive experimental comparison as the aforementioned papers, but instead to provide a detailed understanding of the application of anomaly detection on sleep data and show the challenges of detecting anomalies in this type of human bio signals.

2.2 Identifying mouth breathing

The literature on automatic identification of mouth breathing during sleep is scarce. In the existing literature, mouth breathing is typically identified with questionnaires (Sano et al 2018) and direct observation (de Castilho et al 2016). Mouth breathing measurement is still not commonly included in a standard polysomnography recording and moreover, not manually labelled as a standard practice. One reason is the lack of a reliable and non-invasive measurement device. An oronasal cannula can separate the airflow, but is easily misplaced or fully removed during sleep. An orosonasal thermistor captures the oral flow by measuring the temperature above the mouth (Koutsourelakis et al 2006), but thermistors have a low signal quality (Sabil et al 2019). A specialized mask can be used to separate the breathing channels, but wearing the device may bias the breathing and it is not suitable for children (Hudgel et al 1984). Curran et al. differentiated between the breathing channels by processing the sound during sleep (Curran et al 2012). They applied a Fast Fourier transformation to the raw audio signal, split it into windows of of 5–15 s and trained a deep neural network on this input. Their work is only comparable to a certain extend, because their data stemmed from recordings during awake in a controlled environment where the participants were instructed to breath through their mouth or their nose with an airflow of 1.7 l per s. The data used in our research reflects the real-world conditions of noisy signals and uncontrolled airflow during sleep.

3 Method

3.1 Data

This paper is based on a a comprehensive data set[1] that includes paediatric sleep recordings, including 10–13-year-old children with parent reported sleep-disordered breathing symptoms and a gender and age-matched control group. The study cohort consists of 111 children from the Icelandic EuroPrevall-iFAAM birth cohort research (Grabenhenrich et al 2020; Keil et al 2010; Sigurdardóttir et al 2021) conducted at the Landspitali University Hospital. The sleep recording was done with a Nox Medical A1 polysomnography (PSG) device.

[1] The access to this data was granted by the National Bioethics Committee of Iceland. The study was approved by the Data Protection Agency of Iceland and includes written consent from each child's legal guardian. We cannot make the data publicly available, as is it protected by the ethical approval.

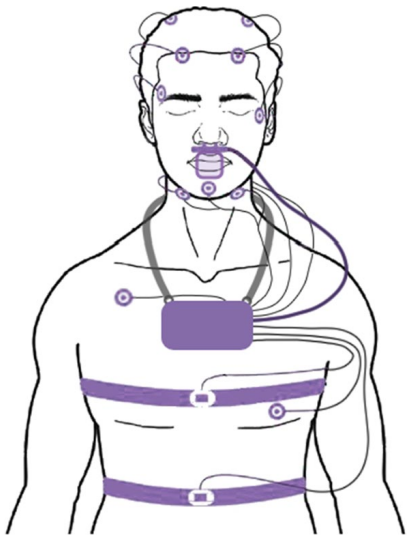


Fig. 1 A polysomnography set-up with an oronasal cannula

PSG is the continuous recording of physiologic activity during sleep. The measurement includes the following sensors: Electroencephalogram (EEG), electrooculography (EOG), chin and leg electromyography (EMG), electrocardiography (ECG), pulse oximeter, microphone, electrodermal activity (EDA) sensor, and accelerometry measuring the movement and body position. Thoracic and abdominal respiratory inductance plethysmography (RIP) belts measure the inflation and deflation of the chest during breathing and an oronasal cannula with separate pressure outputs, monitors the nasal and oral airflow, respectively (Markun and Sampat 2020). A visualization of a PSG set up with an oronasal cannula can be seen in Fig. 1. The PureFlow oronasal cannula by Braebon is specially designed to capture the oral flow and nasal flow separately and was utilized in this study, but is not included in a standard PSG. The PSG was set up at the hospital by sleep technologists, but the participants slept at home and returned the devices the next morning (Kainulainen et al 2021).

Each PSG recording was approximately 8 h long, containing 84 different signals in total. We focused on the nasal and oral flow as well as the thorax and abdomen movement, which measured breathing or movement. Additionally, we included blood oxygen saturation, the audio volume, the heart rate, and body position in the analysis. Two exemplary sequences of the respiratory signals can be seen in Fig. 2, where the thorax movement is colored in light blue, the abdomen movement dark blue, the nasal flow dark green, and the oral flow light green. The top shows a typical nose breathing sequence and the bottom shows a typical mouth breathing sequence. In the mouth breathing sequence, the oral flow shows higher amplitudes than the nasal flow and the amplitudes of the nasal flow are lower than during nose breathing. This behavior is typical for mouth breathing, but cannot be generalized to all mouth breathing sequences. There were 111

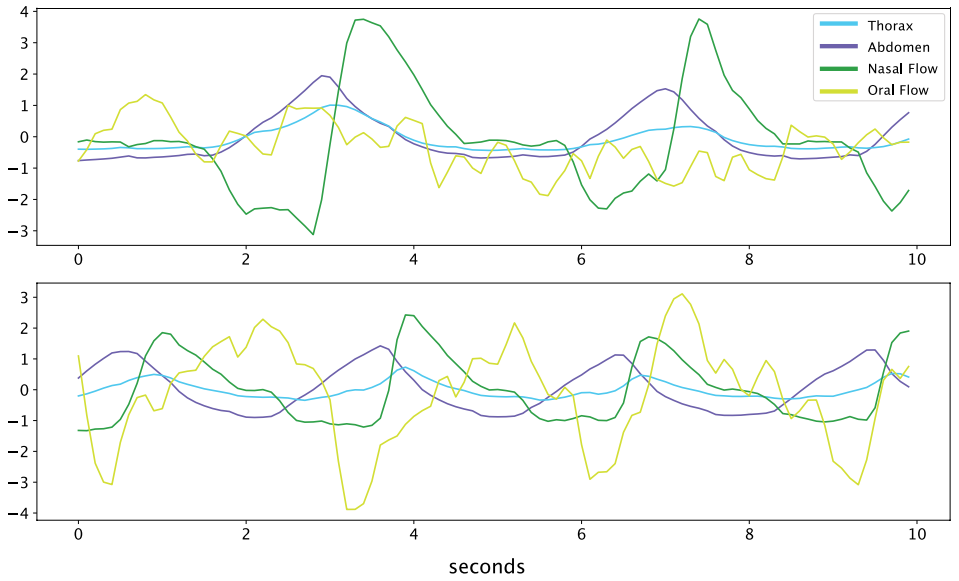


Fig. 2 Two exemplary breathing periods of 10 s each. The top signal is a nose breathing sequence and the bottom is a mouth breathing sequence

recordings, of which 20 have been manually labeled by a sleep technologist. The manual labels were based on the oral and nasal flow signals. 10 recordings were manually chosen to have 5 healthy children and 5 children with sleep disordered breathing. The latter 10 recordings were chosen because the parent-reported information from the questionnaires indicated mouth breathing.

3.2 Preprocessing

The data of PSG recordings were saved in the.edf standard format, an open-source file format commonly used for medical data in Europe. It is designed for multi-channel medical time series and allows different sampling frequencies for each signal (Kemp et al 1992). For each PSG, we extracted the signals of interest and each signal's sampling frequency. Some PSGs had faulty or missing recordings from the RIP belts or the cannula, therefore each recording was visually checked for completeness. During this process, two labeled studies were removed due to low signal quality or measurement errors, which led to a total of 18 eligible labeled sleep studies for this paper.

The four respiratory signals have a sampling frequency of 200 Hz. As the average duration of one study is 8 h, one PSG contains on average 5,760,000 values per signal. Therefore, the full data set quickly becomes complex. To process this large amount of data, we faced a trade-off between run-time and the completeness of the data representation. We downsampled the signals to a sampling frequency of 10 Hz, because this reduced the complexity, but still captured the relevant

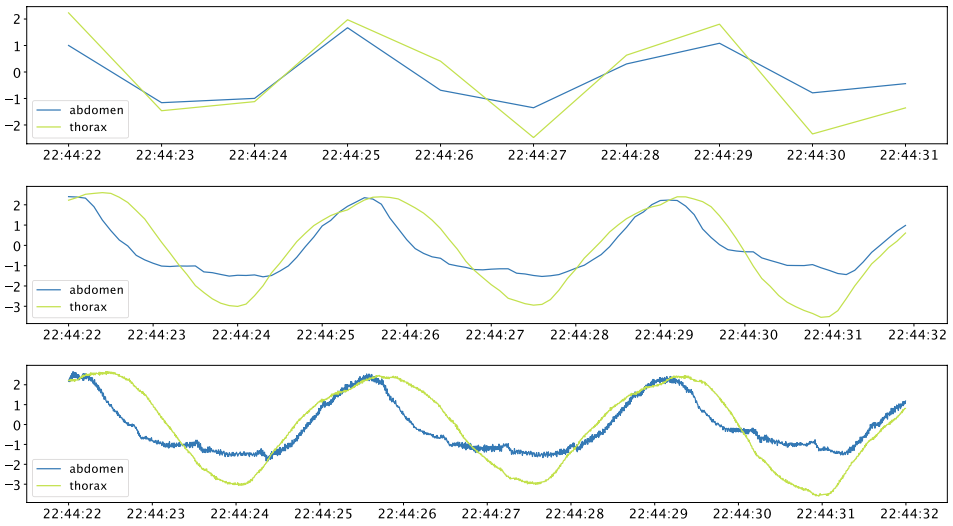


Fig. 3 The same 10 s interval in 1 Hz at the top, 10 Hz in the middle, and 200 Hz (original sampling frequency) at the bottom

features in the data. We furthermore upsampled the oxygen saturation and heart rate with a sampling frequency of 3 Hz to have a common sample rate (Gao et al 2018). The differences between too simplified downsampling to 1 Hz, the chosen downsampling to 10 Hz, and the original sampling frequency of 200 Hz can be seen in Fig. 3.

The signals have different scales, as the oral flow has a range from approximately −2 cm H_2O to 1 cm H_2O, while the thorax and abdomen only range between − 0.0005 V and 0.0005 V. As some models are sensitive to different sized scales, we prevented the signals with larger scales to out rule the signals with smaller scales, by scaling the data to the same range using the scikit-learn StandardScaler. In order to treat this data set as a time series classification problem, we split the data into sequences of 10 s. One respiration cycle, i.e., inhalation and exhalation, of children that are 6 years and older usually has a duration of 2–5 s during sleep (Fleming et al 2011). Choosing an interval of 10 s guarantees that the interval contains at least one full breath and up to 5 full breaths. We did not choose a longer duration than 10 s, to do the classification as granular as possible and keep the complexity of the data, i.e., the length of the time series, as low as possible. We tested both disjoint splits and sliding window splits but as no visible difference in model performance was perceived, we chose the less complex method of disjoint splits. The target variable 'breathing channel' was labeled as 1 for mouth breathing or 0 for nose breathing. It was assigned to each sequence based on the annotation file. To be considered as target class 1, the sequence had to contain at least 3 s (the average length of one respiration cycle) of mouth breathing according to the manual labels by the sleep technologist.

3.3 Model training and hyperparameter optimization

The recording of each participant included approximately 2500 sequences. Since each participant has individual characteristics, their sleep recordings have individual characteristics as well, which is why we used a leave-one-out cross validation for model training and evaluation, i.e., we trained the model on all participants but one and tested the model performance on the test individual. This way, we ensured that no data of the test individual leaked into the model training. For the hyperparameter optimization, a validation set of 3 participants was separated from the rest of the participants. These recordings were only used for hyperparameter optimization and were not included in the leave-one-out cross validation. The validation participants included one with more than 10 mouth breathing sequences, one with zero mouth breathing sequences and one with low signal quality to represent different types of recordings that were present in the data set. We optimized the hyperparameters of all deep learning models on this separate validation set with a keras RandomSearch. For this we defined a grid of possible values for the number of filters, the kernel size, the dropout rate and the learning rate, from which the RandomSearch randomly selected parameter combinations. The hyperparameter optimization has been conducted in the same method and same extend for all deep learning models. The unsupervised deep learning models were optimized to achieve a maximal accuracy in the validation set. The autoencoder was optimized with a a custom loss function to to maximize the average reconstruction error of mouth breathing divided by the average reconstruction error or nose breathing. The hyperparameters of the feature-based model were optimized with a RandomSearch as well. Here, we tuned the learning rate and number of estimators. The hyperparameters for the time series classifiers were set through testing different values on this validation set manually.

3.4 Machine learning methods

As a comparison of multiple machine learning methods, we compare three different time series classifiers and two supervised deep learning models using the raw time series as an input. Ultimately, we propose a reconstruction-based method and feature-based method. Each of the methods will be described in the following.

3.4.1 Supervised time series and deep learning models

All three time series models work with different representations of the data. This includes using the full sequence, fixed intervals, or dynamic shapelets in the training (Bagnall et al 2017). A brief explanation of each model and the selected parameters can be seen in Table 1. We also test two deep learning models, since they can handle multivariate time series, which allows them to capture the interaction between signals.

Table 1 A description of the supervised time series classifier and deep learning models used for benchmarking

Model	Description of approach and parameters
KNN-DTW	The K-Nearest Neighbour classifier (KNN-DTW) calculates the distance of the full sequence to all other sequences, using distance time warping. Then it uses the label of the k nearest neighbors to classify the sequence (Ratanamahatana and Keogh 2005). We chose k=10 and balance the train set with downsampling
TSF	The Time Series Forest (TSF) splits the sequences into intervals and calculates summary statistics. It only considers the 'important' areas of the sequence. First, one classifier is trained for each signal, then all classifiers are combined as a Time Series Forest Ensemble. We chose an ensemble size of 500 and balance the train set with downsampling (Deng et al 2013)
MRSEQL	The Multiple Representation Sequence Learner (MRSEQL) transforms each sequence into a symbolic representation and selects discriminative subsequences, shapelets, for the classification (Le Nguyen et al 2019). We chose both the Symbolic Aggregate Approximation and the Symbolic Fourier Transformation
RNN	The Recurrent Neural Network (RNN) can capture temporal dependencies and complex non-linear correlations within the data by using long short term memory (LSTM) layers (Malhotra et al 2015). We created a model with an LSTM layer of size 100, a dropout layer with a dropout rate of 0.2, and a dense output layer
CNN	A Convolutional Neural Network (CNN) transforms the time series data with convolutional filters and MaxPooling operations (Zhao et al 2017). We created a model with two hidden layers. The first convolutional layer had 64 filters of size 1 and is followed by a MaxPooling and a Dropout layer with a dropout rate of 0.2. The second convolutional layer had 16 filters with a size of 10. This layer was again proceeded by a MaxPooling layer and a Dropout layer. Finally, a Dense layer did the classification

3.4.2 Reconstruction-based anomaly detection

Autoencoders are commonly implemented with multi-layer neural networks. They learn an encoding and a decoding function using an iterative optimization process. The data is passed through the network, the reconstruction error is calculated and at each iteration, the weights of the network are updated (LeCun et al 2015). In a convolutional autoencoder, convolutional layers are included in the encoder and deconvolutional layers in the decoder of the neural network (Ribeiro et al 2018). Convolutional layers transform the data by sliding a filter over the time series. Several filters of different sizes can be applied to learn multiple discriminative features from the input time series. The deconvolutional layers, or transposed convolutions, work by the same principle but swap the forward and backward passes of the convolution. Average- or MaxPooling reduces the length of a time series by aggregating it with a sliding window (Fawaz et al 2019). The hidden layers aim to separate relevant and irrelevant features, which can hide the presence of anomalies (Chalapathy and Chawla 2019). In the encoder, lowering the dimensionality of the input with the convolutional layers, creates a bottleneck after which ideally only the most explanatory parts of the data remain. In the decoder, the transposed convolutions increase the dimensionality of the data back into its original shape. The new representation, i.e the reconstructed input, will naturally differ from the original representation. However, this deviation is encouraged since it did not happen at random, but is a result of

the learned weights of the autoencoder. Autoencoders are trained with the objective of minimizing the reconstruction error, i.e., the error between the original input and the reconstructed output (Li et al 2020). As the majority of the examples in the training data belong to the normal class, the autoencoder will mainly learn the properties of this normal class (Chalapathy and Chawla 2019). Reconstruction-based anomaly detection relies on the assumption that the reconstruction of anomalies is less accurate than the reconstruction of normal instances. As a result, the reconstruction error is higher for anomalous examples, which allows us to use it as an anomaly score and detect anomalies in a fully unsupervised way (Chandola et al 2009).

We used the Root Mean Squared Error (RMSE) as distance metric for defining the reconstruction error, because the input and output of our model were multivariate time series. We chose RMSE above other distance metrics as it gives relatively high weight to large errors. For each 10-second interval, we calculate the RSME by averaging the squared root of the distance between the original and reconstructed signal at all 100 time steps. Equation 1 shows how the RMSE for signal j is calculated over all $n{=}100$ time steps:

$$RMSE_j = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2}{n}} \tag{1}$$

In the model training, we used the average RMSE of all included signals as the loss function for optimization. In the reconstruction-based anomaly detection, we used the RMSEs of the individual signals as new features. In order to use these new features for reconstruction-based anomaly detection, we need to define a classification threshold t. All examples with a higher reconstruction error than t are classified as anomalies and all examples with a lower reconstruction error than t were classified as normal instances. Hence, we classify all examples above t as mouth breathing, and all examples below t as nose breathing. Defining an appropriate threshold is crucial for the success of the anomaly detection. We ran experiments with different approaches of setting the threshold to find the one which achieves the most accurate classification. The most straightforward approach is taking the average reconstruction error of all signals. We also took the distribution of the data into account by adding the standard deviation of the reconstruction error to the threshold. Thus, we defined the threshold t as the average reconstruction error plus the average standard deviation over s signals as shown in Eq. 2.

$$t = \frac{1}{s} * \sum_{i=1}^{s} RMSE_i + \frac{1}{s} * \sum_{i=1}^{s} \sigma_i \tag{2}$$

Knowing the reconstruction error of each signal individually, allows us to use subsets of the signals for defining a threshold as well. Figure 4, a correlation matrix between the reconstruction errors and the breathing channel, here referred to as target variable y, shows that some reconstruction errors correlate more with the breathing channel than others. The highest correlation can be seen between y and the oral pressure. For this reason, we propose a second approach of setting the threshold,

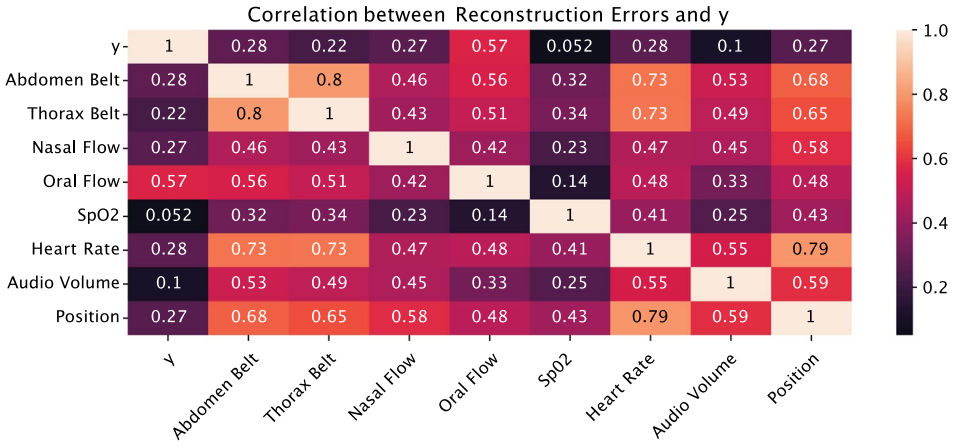


Fig. 4 Correlation matrix between the reconstruction errors and the target variable y

using only the most discriminative feature, the oral flow, instead of the average. The definition of this threshold t is shown in Eq. 3.

$$t = RMSE_{oral} + \sigma_{oral} \tag{3}$$

During this experimenting, we observe that sequences with bad signal quality have extremely high overall average reconstruction errors. Hence, we define a threshold that uses both the reconstruction error of the oral flow and the average reconstruction error. We extend the definition from Eq. 3 by adding a constraint on the average reconstruction. Now, we furthermore discard sequences which have a higher average reconstruction error than 99% of all other sequences.

The autoencoder in this paper was implemented with a convolutional neural network using TensorFlow and is built as a keras sequential model. We constructed the autoencoder by starting with a simple set-up, the input layer, one convolutional layer, and one transposed deconvolutional layer followed by a max-pooling- or respectively upsampling layer with all default values. To avoid the risk of overfitting, we added dropout layers for regularization. Then the complexity of the model was gradually increased and an autoencoder with two hidden layers in the encoder and decoder was chosen. One important property of the model is the dimensionality of the latent space. The representation of the data in this state is crucial for the reconstruction and therefore the success of the anomaly detection. A too big latent space prevents the data from learning a model at all. In the most extreme case, with a latent space of the same size as the input space, the reconstruction error is zero and no classification is possible. Choosing a too small latent space is also not recommended, as too much information is lost in the bottleneck. We choose a range of possible filter and kernel sizes that do not allow a too small or too big latent space in the hyperparameter optimization. The full autoencoder architecture can be found in Fig. 9 in the appendix. The hyperparameter tuning results in an optimal learning rate of 0.001, using the RMSProp optimizer. Finally, we train the model in 50 epochs with a batch size of 256.

We trained the unsupervised model on a bigger train set than the other models, as it additionally included unlabeled recordings, but we tested it on the same test set as the other models. Since the participants of the study were not randomly selected, but consisted of 50% of children with a history of sleep-disordered breathing and 50% of a control group with a history of normal breathing, we did a pre-selection for the train set. This step aimed to lower the number of mouth breathing sequences in the train set to a level that reflects the average population better. This pre-selection was done based on a parental questionnaire regarding the child's breathing behavior during sleep. If the parents answered that they observed their children sleeping with an open mouth, waking up with a dry mouth, or breathing through their mouth during the day, we disqualified the child from the train set. This led to disqualifying 54 children from the training, which approximately reflects the percentage of the study population with abnormal breathing behavior. Including these children in the train set could contradict the assumption that mouth breathing is the rare exception. It is still possible, that mouth breathing was included in these recordings, but we can assume that the proportion of mouth breathing in this subset of recordings was low.

Reconstruction-based anomaly detection can also be implemented as a semi-supervised model. Similar to the unsupervised approach, the autoencoder is trained without using any labels and no labeled anomalous examples are needed. Instead, we use only examples of the normal class for the model training as described in Chalapathy and Chawla (2019). The idea behind a semi-supervised approach is to train the autoencoder only on sequences that certainly do not contain any mouth breathing. This way, we do not have to rely on the assumption that the imbalance in the data set is high enough to disregard the mouth breathing sequences in the training data.

3.4.3 Feature-based classification

As a comparison to the unsupervised deep learning model, we train a classifier which works with simple statistical features. We transform the 3-dimensional time series data set into a 2-dimensional data set with time-independent features. This is done by calculating summary values for each sequence of 100 time steps, including the mean, standard deviation, minimum and maximum of each signal. Additionally, we create two more features based on the oronasal cannula. We calculate the difference between the mean of the oral flow and nasal flow, as well as the difference between the standard deviation of the oral flow and nasal flow, which can be seen in Eq. 4, where n is the number of time steps.

$$\text{Oronasal Difference} = \frac{1}{n} * \sum_{i=1}^{n} \text{Oral Flow} - \frac{1}{n} * \sum_{i=1}^{n} \text{Nasal Flow} \tag{4}$$

We then applied a feature selection based on the Pearson correlation coefficient between the feature and the target variable. We selected the 10 most correlated features for the model training. Figure 5 shows the correlations of the 10 selected features.

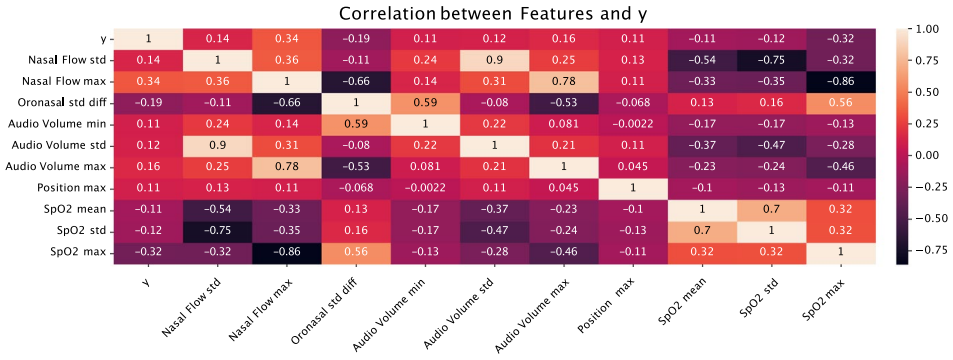


Fig. 5 Correlation of features with target variable mouth breathing

These features were then used as an input for the supervised classification method gradient boosting machine (GBM). A GBM is an ensemble method that commonly excels over other machine learning methods in bench marking studies or practical applications (Natekin and Knoll 2013). Ensemble methods use the combined classification power of multiple individual machine learning classifiers. While other ensemble methods simply average the predictions of multiple classifiers, the strength of a GBM lies in the sequential training of models, which allows it to take the classification errors of the previous models into account. We trained the GBM with a learning rate of 1 and use 1000 estimators.

3.5 Evaluation

We evaluated the models by comparing the predicted labels by our models to the manual labels provided by the sleep technologist. To achieve a comprehensive evaluation of how the models can predict mouth breathing in unseen sleep recordings, we performed a leave-one-out cross validation. This form of cross validation evaluates the models for each sleep recording individually by using the recordings of all children but one for training and the remaining one for testing. This way, we could observe the inter-subject variability and use a higher amount of training data. We calculated the average precision and recall scores as proposed by Forman and Scholz (2010) to avoid bias from the class imbalance in different folds. As we are facing a highly imbalanced classification problem, it does not make sense to consider accuracy as an evaluation measure. Instead, we rely on metrics, which evaluate the classification of the minority class, such as precision, recall, and F1 score. For the final results, we added up the confusion matrices of all folds for each model and calculated the precision, recall and F1 score from the total number of true positives, false positives and false negatives. The individual classification results of the models on each child's recording can be found in Table 4 in the appendix.

We furthermore calculated the standard deviation of these different metrics across the participants. This showed how much the classification performance varies across participants and therefore how good the model can generalize on test sets with unique characteristics. Another perspective towards the anomaly detection is

added when we divide the participants in the test set by a high or low number of mouth breathing, which represents the group of healthy and sleep disordered participants. This analysis reveals which models can handle test sets with close to zero positive examples. In this group we additionally calculate the False Positive Rate (FPR) by dividing the number of false positives by the number of false positives and true negatives. This test shows how strongly a model would overestimate the degree of mouth breathing in an healthy individual.

4 Results

4.1 Naive baseline

As a naive baseline, we do stratified random guessing, which takes the distribution of nose breathing and mouth breathing examples in the train set into account. Each sample in the test set gets the label nose breathing or mouth breathing with a probability that reflects the class distribution. This approach resulted in an F1 score lower than 0.01, which gives a hint at the difficulty of classification in a highly imbalanced data set.

4.2 Overall evaluation

All supervised models were evaluated within the same leave-one-out cross validation as the reconstruction-based anomaly detection to achieve comparability between all models. Table 2 shows the performance of all models evaluated on 15 different test folds using the leave-one-out cross validation. Evaluating the overall performance of the machine learning models, showed that the GBM using statistical features as an input is the best classifier with an F1 score of 0.54. The reconstruction-based classifier has a similar F1 score of 0.508. Comparing the classification accuracy of all supervised models showed, that the deep learning models were not

Table 2 Comparison of average classification accuracy, standard deviation among all folds in brackets and training time for time series classifiers, supervised deep learning models, and autoencoders

Classifier	Type	Training time	Precision	Recall	F1 score
Random	Naive baseline	–	0.021	0.022	0.022
GBM	Feature-based	2 min	**0.445**	**0.704**	**0.546**
KNN-DTW	Similarity-based	15 min	**0.256**	0.477	0.333
TSF	Interval-based	6 min	0.121	0.243	0.162
MRSEQL	Shapelet-based	31 min	0.231	**0.861**	**0.364**
RNN	Deep learning	40 min	**0.454**	**0.229**	**0.304**
CNN	Deep learning	4 min	0.350	0.120	0.179
Autoencoder	Reconstruction-based	8 min	**0.401**	**0.695**	**0.508**

The best-performing method for each approach is shown in bold

necessarily better than the time series classifiers. The classifiers KNN-DTW and MRSEQL, which can handle multivariate time series as an input, performed better than the TSF, which combines the predictions of the individual time series in an ensemble. The best performing supervised model was MRSEQL, which even exceeded the performance of the deep learning models. Both deep learning models have a lower classification accuracy than the classic feature-based classifier. The recall, i.e., how many of the true mouth breathing sequences were identified as such, and the precision, i.e., how many of the predicted mouth breathing were correct, give an enhanced insight on the model performance. Most machine learning models in the evaluation had a high recall but a low precision. This means they identified many true positives, but also predicted many false positives. The only exceptions to this trend were the deep learning models. Both the RNN and the CNN had a higher precision than recall. Especially the CNN led to a low recall score, as it fails to identify most of the true mouth breathing sequences. Looking at the individual training folds showed, that for some participants the CNN was not able to make any prediction and resulted in an F1 score of 0.

4.3 Individual-level evaluation

Evaluating the performance of the machine learning models on an individual level showed that the classification accuracy of all models varies strongly. The best performing models, the feature-based classifier and the reconstruction-based classifier both had a standard deviation of the F1 score of 0.3. This shows that the models work well for some participants and perform poorly on other participants. Reviewing the participants one by one showed that the performance of all models was weaker for the participants with a lower amount of mouth breathing. This is plausible, as an increased imbalance ratio affects the classifier performance as shown by Lemnaru and Potolea (2011). Furthermore, having evaluation folds with zero positive examples in the test set can naturally only lead to a decrease of precision, recall and F1 score as achieving true positive classifications is impossible in this setting. However, it is a relevant results since these participants with zero or few mouth breathing sequences represent healthy, non-mouth breathing participants, which is

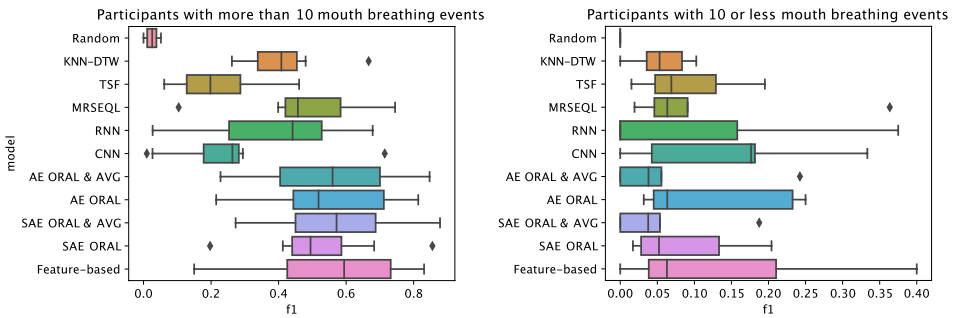


Fig. 6 Distribution of F1 score across participants with a low number of mouth breathing ($n = 8$) and a high number of mouth breathing ($n = 7$)

the majority of the population. Dividing the test set into participants with more than 10 mouth breathing sequences and lower or equal to 10 mouth breathing sequences shows how each model would perform when classifying healthy or sleep disordered participants.

Figure 6 shows the F1 score of the individual participants as a distribution for each machine learning model. The plot on the left side shows the low or non-mouth breathing participants and the plot on the right side shows the participants with more than 10 mouth breathing sequences. We can see that the CNN, which did not perform well in the overall evaluation with an F1 score of 0.179 was the best performing model for the participants with a low-amount of mouth breathing. The reconstruction-based model and the feature-based model both had a low performance on the low-mouth breathing sequences. However, the autoencoders and the GBM were leading the performance in the high-mouth breathing participants. Both the CNN and the RNN have a False Positive Rate (FPR) of 0.007. This equals to approximately 15 false positives on average in each participant. Even though the feature-based classifier in comparison has a FPR of 0.015 with 33 false positives on average.

4.4 Reconstruction-based anomaly detection

Applying the autoencoder on an unseen test set resulted in a reconstruction error that was indeed higher for mouth breathing than for nose breathing. The average reconstruction error of the anomalous class was twice as high as the average reconstruction error of the normal class. This can be seen in Fig. 7, which shows the distribution of the reconstruction errors of mouth breathing and nose breathing separately.

We can see that most of the nose breathing examples (92.8%) have a reconstruction error below 1. The mouth breathing examples have higher reconstruction errors in a range between 0.5 and 5. This shows different distributions of the reconstruction

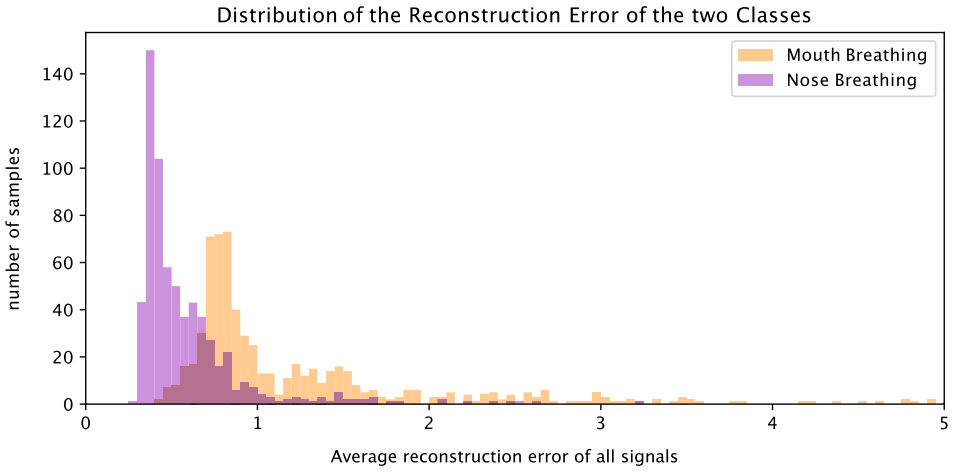


Fig. 7 The distribution of the reconstruction error by target class (nose breathing in purple, mouth breathing in orange). For visualization, the majority class is downsampled to the size of the minority class

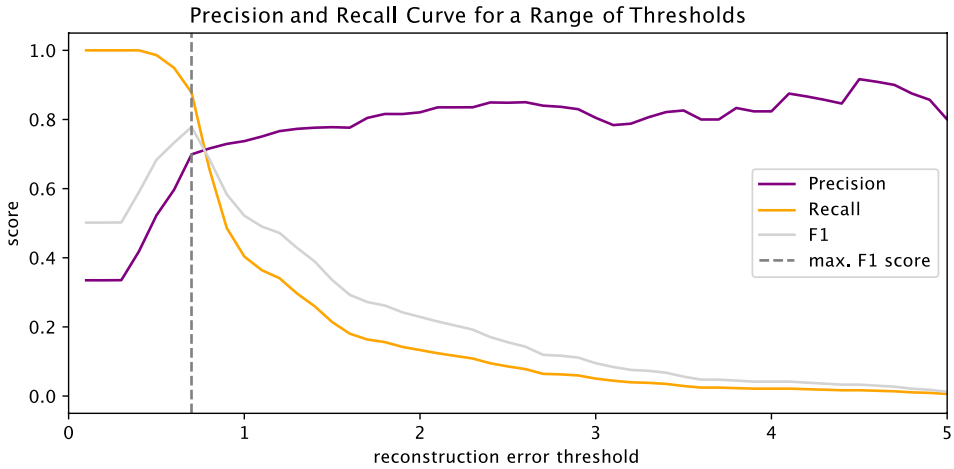


Fig. 8 Precision and recall for the classification at different threshold values

Table 3 Classification accuracy for unsupervised and semi-supervised models

Autoencoder	Signals used for the threshold	Precision	Recall	F1
Unsupervised	Average of all signals	0.107	0.380	0.167
	Oral flow	0.308	**0.721**	0.431
	Average signals & oral flow	**0.352**	0.659	**0.459**
Semi-supervised	Average of all ssgnals	0.094	0.383	0.151
	Oral Flow	0.243	**0.827**	0.376
	Average signals & oral flow	**0.401**	0.695	**0.508**

The highest value for each approach is shown in bold

error of the two classes, which is why we can use the reconstruction error as an anomaly score. However, as the classes do overlap (shown in red in Fig. 7), a perfect separation of normal and anomalous data by threshold was impossible based on the reconstruction error.

Figure 8 shows how precision, recall and F1 score changed by gradually increasing the threshold value. Testing the classification accuracy of the different threshold approaches showed that the way the threshold was defined had a high impact on the classification performance. The estimated thresholds used for classification did not necessarily match the optimal thresholds like the one shown in Fig. 8 as the grey dotted line. Figure 8 shows that even small deviations from this optimal separation led to a strong decrease in classification performance. Hence, the following results show the classification ability of this particular unsupervised classification approach, but do not necessarily reflect the full potential of the autoencoder for reconstruction-based anomaly detection.

The summarized results of evaluating the unsupervised- and semi-supervised autoencoder are shown in Table 3. The classification with the average threshold led to low results even though theoretically a separation of the classes is given as seen

in Fig. 7. The average threshold is not only higher for mouth breathing but also in bad signal quality. For this reason, it was not suitable for the unsupervised detection of mouth breathing. The results strongly improved when only the oral threshold is used for the classification. This approach achieved the highest recall of 0.827 in the semi-supervised training. We could further improve the F1 score of this approach by combining the information from the average and mouth breathing reconstruction error. This approach achieved the best overall results in the semi-supervised training with a precision of 0.401, a recall of 0.695 and an F1 score of 0.508.

4.5 Error analysis

In order to gain a deeper understanding of the classification performance, we review a subset of the misclassified sequences of the reconstruction based anomaly detection with a sleep technologist. In particular, we review the false positives, i.e the nose breathing sequences the model labels as mouth breathing. We take the time stamps of a subsample of the test set and review these sequences in the sleep analysis software Noxturnal by Noxmedical. The following reasons for misclassifications were identified:

- **Slight or short mouth breathing:** Some sequences show mouth breathing in the manual review but were not labeled as such, because it was only slight mouth breathing. In many of these cases, it was shortly before or after a labeled sequence of mouth breathing. Others were correctly labeled but were disregarded in the preprocessing because the mouth breathing was very short and only sequences with at least three seconds of mouth breathing were considered as mouth breathing sequences.
- **Mouth breathing during awake state:** There are multiple sequences that actually had mouth breathing but were not labeled as such by the sleep technologist, because they occurred during an awakening, which is not considered as clinically relevant.
- **Bad signal quality:** In some sequences, the sleep technologist cannot decide whether mouth breathing is present, because the signals are noisy or include artifacts. In one sequence it is clearly visible that the oximeter lost contact. In other cases, we assume that the oronasal cannula has moved.

This review shows us that the we cannot always rely on the manual labels as the ground truth. Setting clear borders where mouth breathing starts and stops is a challenge for the human reviewer as well, especially in recordings with bad signal quality. We should keep in mind that assigning the labels is a subjective task and that interrater variability is an ongoing research area in sleep (Danker-hopfe et al 2009). It also reveals a weakness of the preprocessing, which indicates we should lower the required minimum amount of mouth breathing per sequence in future work. Whether we should exclude mouth breathing during awakenings from the evaluation or include information about the sleep stages in the input data remains an open question. Overall, the error analysis also showed that many of the false positives have not been entirely false after all.

5 Discussion

The results demonstrated that machine learning can be used to automatically differentiate between mouth breathing and nose breathing. The comparison of time series classifiers, deep learning models, unsupervised models and a feature-based classifier showed that overall the feature-based classifier was the best performing machine learning method. Evaluating the performance of these models in the leave-one-out cross validation showed that the model performance varied strongly across participants. The two best performing models, the reconstruction-based anomaly detection and the feature-based classifier showed a similar standard deviation. They also showed a similar performance drop on the test set with a low number of mouth breathing in comparison to the test set with participants with a high number of mouth breathing. This showed that both models may overestimate the severity of mouth breathing when used in clinical practice.

To assess whether this classification accuracy is precise enough to replace manual annotation work, we should consider the implications of false positive and false negative classifications for the sleep technologists, as well as the consequences that arise for the child. Classifying too many nose breathing sequences as mouth breathing sequences gives the impression that a child suffers from a condition they do not have or only mildly suffer from. On the contrary, capturing none or too few of the true mouth breathing sequences may lead to underestimating the severity of mouth breathing and preventing the child from receiving the appropriate diagnosis and treatment. Our best performing model has a precision of 44.5% and can identify 70.4% of all mouth breathing sequences. Therefore, it is likely to identify a high percentage of the mouth breathing sequences but may overestimate the mouth breathing. Both the feature-based and reconstruction-based methods have a low precision but high recall. Therefore they could be suitable to support the sleep technologist by highlighting the sequences which are likely to be mouth breathing and leave the final decision to the expert. This approach of supporting the medical staff instead of fully replacing medical staff has shown success when integrating machine learning applications in clinical practice (Henry et al 2022). Whether sleep technologists rely completely on the prediction in the future or use it as a reference value for faster manual review depends on the desired accuracy of the mouth breathing labels, but either way decreases the manual labeling effort.

Applying the reconstruction-based anomaly detection approach on sleep data and observing separation of the classes by reconstruction error shows that this approach is applicable to sleep data. We can see that the unsupervised approach has a lower classification accuracy than the semi-supervised approach. There are several reasons which may account for this gap. Firstly, the remaining mouth breathing sequences in the train set of the unsupervised approach negatively impact the reconstruction-based anomaly detection. This would show that our proposed model strongly relies on the assumption of an imbalanced data set. Secondly, we also include non-labeled recordings in the train set of the unsupervised model. As these have not been reviewed manually, we have no information of the amount of mouth breathing or the signal quality in these recordings. Another limitation of the reconstruction-based anomaly detection is that

the autoencoder is not able to differentiate between different types of anomalies. Even though we assume that mouth breathing sequences are anomalies, we cannot assume that all anomalies are mouth breathing. Consequently, the false positives, that are incorrectly classified as mouth breathing partly also point towards other anomalies such as measurement errors, which makes our model less applicable for low quality recordings.

However, the comparison of different methods showed that a classic machine learning approach outperforms the deep learning models. The feature-based classifier simplifies the time series into summary features. This shows that the shape of the signals is not relevant for the classification, but rather their altitude and range. Therefore, our research shows, that for this specific application, deep learning models are not superior to classic machine learning models. This goes in line with previous publications questioning the need for deep learning in other domains (Gunnarsson et al 2021; Shwartz-Ziv and Armon 2022). It is an ongoing debate when and how deep learning is needed. The superior performance of our model in comparison to the model by Curran et al (2012) may arise from including more features than only the audio signal. However, including more signals has not only advantages, as a PSG study is more of an effort than a microphone study. The overall results show, that the signals we included in the model are suitable for identifying mouth breathing. It is surprising that the statistical features and the reconstruction error show different correlation with the target variable mouth breathing, as shown in Figs. 4 and 5. While the reconstruction error of the autoencoder mainly shows correlation of the oral flow and the target variable, the statistical features also shows correlations to the audio volume and the oxygen saturation. One reason for that could be that these signals have different properties and are the blood oxygen and audio volume are more meaningful as summary statistics and the oral flow signal is more meaningful as a raw signal. A mixed approach of inputting the oral flow as a raw signal and the audio and blood oxygen saturation as statistical features could be an interesting approach to pursue in future work.

We need to keep in mind, that machine learning models can identify complex patterns from the training data but do not have human reasoning. In one sleep recording, the oronasal cannula is misplaced in a way that the pressure transducer, which captures the mouth breathing, was placed above the nose. For 6 h, the nose breathing signal is gone, but the mouth breathing is unusually high. Listening to the audio and looking at the unusual patterns of the mouth breathing signal, let the sleep technologist conclude that it was a measurement error, even though it looked like extreme mouth breathing. This is a line of thought that comes naturally to the sleep technologist, but can not be achieved by a machine learning model. For this reason, the ability of our machine learning models is limited by the quality of the sleep recording and can be negatively impacted by measurement errors.

6 Conclusion

These findings are relevant for research focusing on sleep-disordered breathing, because they show that mouth breathing can be automatically identified. Using the signals from the oronasal cannula, thorax and abdomen belts, pulse oximeter, and microphone, our proposed approach can classify mouth breathing with an F1 score

of 0.546. This means, that the manual annotation work can be decreased with the use of machine learning. Comparing classic and deep machine learning models, showed that classic methods outperform deep learning and have a higher clinical relevance in this application. The results from the reconstruction-based method showed that we are not dependent on labeled mouth breathing sequences in the training to identify mouth breathing. The results of all machine learning models varied strongly across participants, which highlights the importance of patient-wise evaluation. In future research, we want to test whether these models also work on the data of adults with sleep-disordered breathing. To improve the model performance, the superordinate time series should be taken into account, as a sequence is more likely to be mouth breathing if the preceding and succeeding sequences are mouth breathing as well. The model could be further improved by classifying individual breaths instead of fixed 10-second intervals as proposed in Holm et al (2022).

Appendix 1: Autoencoder architecture

See Fig. 9.

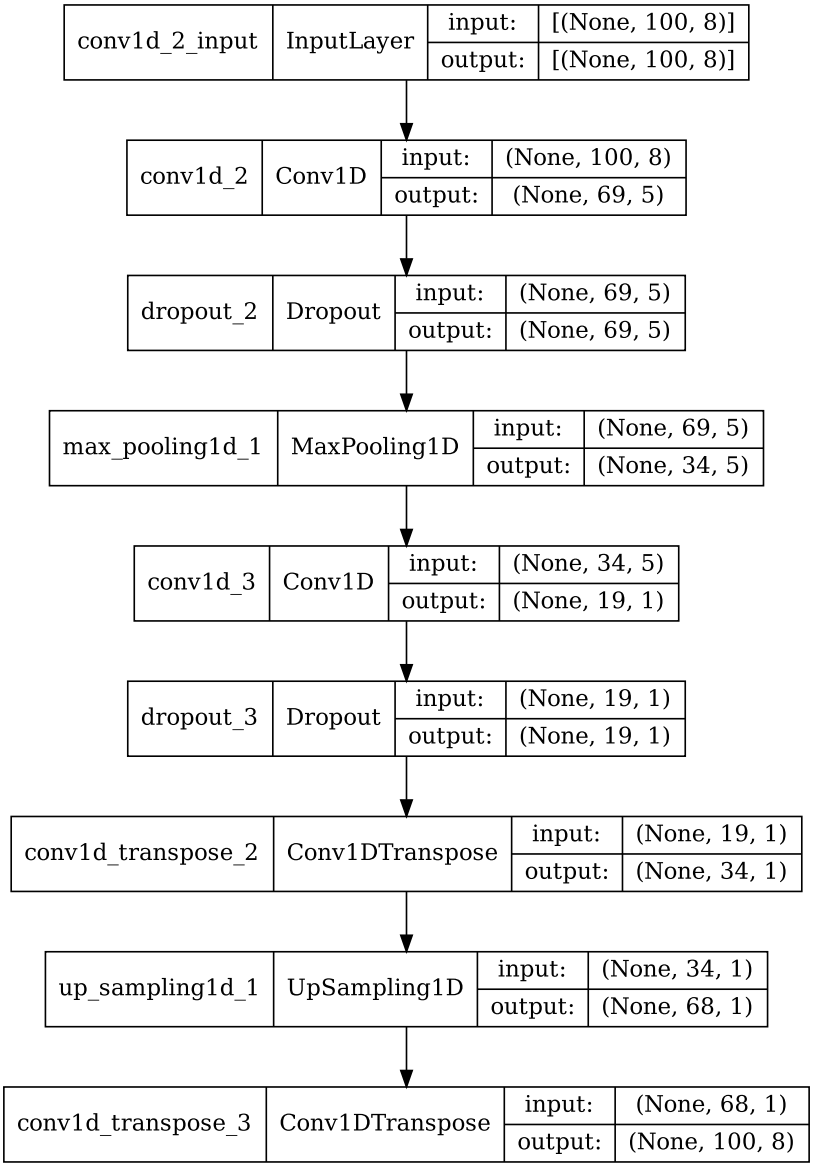


Fig. 9 The architecture of the convolutional autoencoder

Appendix 2: Results of leave-one-out cross validation

See Table 4.

Table 4 Model performance on each test fold

Fold	Model	TN	FP	FN	TP
1	KNN-DTW	1442	29	22	51
	TSF	1423	48	68	5
	MRSEQL	1421	50	0	73
	RNN	1471	0	72	1
	CNN	1470	1	72	1
	AE AVG	1383	88	4	69
	AE ORAL	1439	32	1	72
	AE ORAL & AVG	1446	25	1	72
	SAE AVG	1401	70	17	56
	SAE ORAL	1449	22	2	71
	SAE ORAL & AVG	1452	19	1	72
	GBM	1469	2	27	46
2	KNN-DTW	1206	25	2	0
	TSF	1204	27	0	2
	MRSEQL	1191	40	0	2
	RNN	1228	3	2	0
	CNN	1223	8	1	1
	AE AVG	1161	70	0	2
	AE ORAL	1219	12	0	2
	AE ORAL & AVG	1225	6	2	0
	SAE AVG	1165	66	0	2
	SAE ORAL	1219	12	1	1
	SAE ORAL & AVG	1225	6	2	0
	GBM	1221	10	1	1
3	KNN-DTW	3062	44	30	27
	TSF	3041	65	21	36
	RNN	3102	4	35	22
	MRSEQL	2971	135	4	53
	CNN	3101	5	49	8
	AE AVG	2982	124	17	40
	AE ORAL	3044	62	17	40
	AE ORAL & AVG	3062	44	29	28
	SAE AVG	2979	127	16	41
	SAE ORAL	3020	86	9	48
	SAE ORAL & AVG	3061	45	24	33
	GBM	3088	18	35	22

Table 4 (continued)

Fold	Model	TN	FP	FN	TP
4	KNN-DTW	3284	89	0	0
	TSF	3269	104	0	0
	MRSEQL	3213	160	0	0
	RNN	3372	1	0	0
	CNN	3371	2	0	0
	AE AVG	3221	152	0	0
	AE ORAL	3297	76	0	0
	AE ORAL & AVG	3329	44	0	0
	SAE AVG	3194	179	0	0
	SAE ORAL	3333	40	0	0
	SAE ORAL & AVG	3363	10	0	0
	GBM	3369	4	0	0
5	KNN-DTW	954	32	11	14
	TSF	942	44	16	9
	MRSEQL	930	56	3	22
	RNN	974	12	6	19
	CNN	984	2	10	15
	AE AVG	936	50	16	9
	AE ORAL	956	30	5	20
	AE ORAL & AVG	964	22	8	17
	SAE AVG	927	59	15	10
	SAE ORAL	964	22	7	18
	SAE ORAL & AVG	966	20	6	19
	GBM	956	30	1	24
6	KNN-DTW	1290	9	104	30
	TSF	1276	23	87	47
	MRSEQL	1116	183	7	127
	RNN	1297	2	103	31
	CNN	1299	0	115	19
	AE AVG	1220	79	75	59
	AE ORAL	1284	15	51	83
	AE ORAL & AVG	1289	10	59	75
	SAE AVG	1216	83	74	60
	SAE ORAL	1271	28	50	84
	SAE ORAL & AVG	1281	18	61	73
	GBM	1291	8	74	60

Table 4 (continued)

Fold	Model	TN	FP	FN	TP
7	KNN-DTW	3376	79	0	14
	TSF	3343	112	2	12
	MRSEQL	3214	241	0	14
	RNN	3437	18	8	6
	CNN	3440	15	9	5
	AE AVG	3269	186	0	14
	AE ORAL	3381	74	0	14
	AE ORAL & AVG	3405	50	2	12
	SAE AVG	3259	196	0	14
	SAE ORAL	3341	114	0	14
	SAE ORAL & AVG	3393	62	2	12
	GBM	3383	72	2	12
8	KNN-DTW	2814	82	40	49
	TSF	2779	117	73	16
	MRSEQL	2728	168	25	64
	RNN	2862	34	45	44
	CNN	2886	10	73	16
	AE AVG	2712	184	76	13
	AE ORAL	2786	110	65	24
	AE ORAL & AVG	2814	82	67	22
	SAE AVG	2664	232	71	18
	SAE ORAL	2781	115	36	53
	SAE ORAL & AVG	2830	66	58	31
	GBM	2843	53	36	53
9	KNN-DTW	963	31	4	2
	TSF	963	31	2	4
	MRSEQL	973	21	0	6
	RNN	987	7	3	3
	CNN	990	4	4	2
	AE AVG	950	44	4	2
	AE ORAL	962	32	1	5
	AE ORAL & AVG	971	23	2	4
	SAE AVG	948	46	4	2
	SAE ORAL	956	38	1	5
	SAE ORAL & AVG	971	23	3	3
	GBM	976	18	1	5

Table 4 (continued)

Fold	Model	TN	FP	FN	TP
10	KNN-DTW	1783	52	2	1
	TSF	1755	80	1	2
	MRSEQL	1777	58	1	2
	RNN	1835	0	3	0
	CNN	1834	1	3	0
	AE AVG	1685	150	0	3
	AE ORAL	1776	59	2	1
	AE ORAL & AVG	1790	45	3	0
	SAE AVG	1711	124	0	3
	SAE ORAL	1768	67	2	1
	SAE ORAL & AVG	1787	48	3	0
	GBM	1736	99	1	2
11	KNN-DTW	3348	107	0	3
	TSF	3327	128	2	1
	MRSEQL	3150	305	0	3
	RNN	3423	32	0	3
	CNN	3427	28	0	3
	AE AVG	3250	205	2	1
	AE ORAL	3328	127	0	3
	AE ORAL & AVG	3353	102	0	3
	SAE AVG	2940	515	2	1
	SAE ORAL	3111	344	0	3
	SAE ORAL & AVG	3349	106	0	3
	GBM	3438	17	3	0
12	KNN-DTW	1854	71	0	0
	TSF	1835	90	0	0
	MRSEQL	1869	56	0	0
	RNN	1907	18	0	0
	CNN	1903	22	0	0
	AE AVG	1750	175	0	0
	AE ORAL	1784	141	0	0
	AE ORAL & AVG	1800	125	0	0
	SAE AVG	1795	130	0	0
	SAE ORAL	1855	70	0	0
	SAE ORAL & AVG	1873	52	0	0
	GBM	1911	14	0	0

Table 4 (continued)

Fold	Model	TN	FP	FN	TP
13	KNN-DTW	3312	102	106	96
	TSF	3268	146	191	11
	MRSEQL	3265	149	45	157
	RNN	3410	4	195	7
	CNN	3412	2	201	1
	AE AVG	3236	178	191	11
	AE ORAL	3311	103	34	168
	AE ORAL & AVG	3342	72	39	163
	SAE AVG	3206	208	183	19
	SAE ORAL	2919	495	0	202
	SAE ORAL & AVG	3358	56	29	173
	GBM	3273	141	4	198
14	KNN-DTW	2531	51	15	15
	TSF	2548	34	22	8
	MRSEQL	2530	52	4	26
	RNN	2580	2	19	11
	CNN	2581	1	25	5
	AE AVG	2433	149	12	18
	AE ORAL	2531	51	3	27
	AE ORAL & AVG	2552	30	5	25
	SAE AVG	2429	153	13	17
	SAE ORAL	2528	54	3	27
	SAE ORAL & AVG	2546	36	5	25
	GBM	2574	8	3	27
15	KNN-DTW	2574	88	0	4
	TSF	2582	80	1	3
	MRSEQL	2496	166	0	4
	RNN	2622	40	4	0
	CNN	2620	42	3	1
	AE AVG	2465	197	1	3
	AE ORAL	2544	118	0	4
	AE ORAL & AVG	2563	99	2	2
	SAE AVG	2490	172	1	3
	SAE ORAL	2517	145	0	4
	SAE ORAL & AVG	2562	100	2	2
	GBM	2593	69	2	2

KNN-DTW = K-Nearest Neighbours with Distance Time Warping, TSF = Time Series Forest, MRSEQL = Multiple Representation Sequence Learner, RNN = Recurrent Neural Network, CNN = Convolutional Neural Network, SAE = Semi-supervised Autoencoder, AE= Autoencoder, GBM = Gradient Boosting Machine

Acknowledgements We would like to thank sleep technologist Kristín Anna Ólafsdóttir who was part of the team that set up the PSG devices for the children. Furthermore, we would like to thank the children and their parents for their contribution to this study. This project received funding from the European Unions Horizon 2020 research and innovation program under grant agreement No. 965417. The project was supported financially by the Icelandic Research Fund 2016-2019, No. 174067, Nordforsk 2018-2021, (NordSleep, No. 90458) and the Landspitali University Hospital Science Fund 2019-2020 (No. 893831). Nox Medical (Reykjavik, Iceland) additionally supported the study by supplying the researchers with the A1 sleep recorders and consumables needed for the sleep studies. The birth cohort study was funded by the European Commission: (a) under the 6th Framework Programme (FOOD-CT-2005-514000) within the collaborative research initiative 'EuroPrevall', and (b) under the 7th Framework Programme (FP7-KBBE-2012-6; grant agreement No. 312 147) within the collaborative project 'iFAAM'.

References

Arnardottir ES, Islind AS, Óskarsdóttir M (2021) The future of sleep measurements: a review and perspective. Sleep Med Clin 16(3):447–464

Arnardottir ES, Islind AS, Óskarsdóttir M et al (2022) The sleep revolution project: the concept and objectives. J Sleep Res 31(4):e13,630

Bagnall A, Lines J, Bostrom A et al (2017) The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mini Knowl Discov 31(3):606–660

Biedebach L, Rusanen M, Leppänen T et al (2023) Towards a deeper understanding of sleep stages through their representation in the latent space of variational autoencoders. In: proceedings of the annual Hawaii international conference on system sciences, IEEE Computer Society, pp 3111–3120

Blázquez-García A, Conde A, Mori U et al (2021) A review on outlier/anomaly detection in time series data. ACM Comput Surv (CSUR) 54(3):1–33

Chalapathy R, Chawla S (2019) Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv (CSUR) 41(3):1–58

Chauhan S, Vig L (2015) Anomaly detection in ecg time signals via deep long short-term memory networks. In: 2015 IEEE international conference on data science and advanced analytics (DSAA), IEEE, pp 1–7

Curran K, Yuan P, Coyle D (2012) Using acoustic sensors to discriminate between nasal and mouth breathing. Int J Bioinform Res Appl 8(5–6):382–396

Danker-hopfe H, Anderer P, Zeitlhofer J et al (2009) Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. J Sleep Res 18(1):74–84

de Castilho LS, Abreu MHNG, de Oliveira RB et al (2016) Factors associated with mouth breathing in children with developmental disabilities. Spec Care Dent 36(2):75–79

Deng H, Runger G, Tuv E et al (2013) A time series forest for classification and feature extraction. Inform Sci 239:142–153

Denotti G, Ventura S, Arena O et al (2014) Oral breathing: new early treatment protocol. J Pediat Neonat Individ Med (JPNIM) 3(1):e030,108-e030,108

Fawaz HI, Forestier G, Weber J et al (2019) Deep learning for time series classification: a review. Data Min Knowl Disc 33(4):917–963

Fensterseifer GS, Carpes O, Weckx LLM et al (2013) Mouth breathing in children with learning disorders. Braz J Otorhinolaryngol 79:620–624
Fleming S, Thompson M, Stevens R et al (2011) Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. The Lancet 377(9770):1011–1018
Forman G, Scholz M (2010) Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM Sigkdd Explorat Newsl 12(1):49–57
Freeman C, Merriman J, Beaver I et al (2021) Experimental comparison and survey of twelve time series anomaly detection algorithms. J Artif Intell Res 72:849–899
Fu K, Cheng D, Tu Y, et al (2016) Credit card fraud detection using convolutional neural networks. In: neural information processing: 23rd international conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part III 23, Springer, pp 483–490
Gao J, Murphey YL, Zhu H (2018) Multivariate time series prediction of lane changing behavior using deep neural network. Appl Intell 48(10):3523–3537
Gozal D (1998) Sleep-disordered breathing and school performance in children. Pediatrics 102(3):616–620
Grabenhenrich L, Trendelenburg V, Bellach J et al (2020) Frequency of food allergy in school-aged children in eight European countries-the Europrevall-Ifaam birth cohort. Allergy 75(9):2294–2308
Gunnarsson BR, Vanden Broucke S, Baesens B et al (2021) Deep learning for credit scoring: do or don't? Europ J Operat Res 295(1):292–305
Henry KE, Kornfield R, Sridharan A et al (2022) Human-machine teaming is key to ai adoption: clinicians' experiences with a deployed machine learning system. NPJ Dig Med 5(1):97
Holm B, Óttir M, Arnardóttir ES, et al (2022) Automatic non-invasive isolation of respiratory cycles. arXiv preprint arXiv:2203.01828
Huang G, Ma F (2021) Concad: contrastive learning-based cross attention for sleep apnea detection. In: joint european conference on machine learning and knowledge discovery in databases, Springer, pp 68–84
Hudgel DW, Martin RJ, Johnson B et al (1984) Mechanics of the respiratory system and breathing pattern during sleep in normal humans. J Appl Physiol 56(1):133–137
Izu SC, Itamoto CH, Pradella-Hallinan M et al (2010) Obstructive sleep apnea syndrome (Osas) in mouth breathing children. Braz J Otorhinolaryngol 76:552–556
Kainulainen S, Korkalainen H, Sigurdardóttir S et al (2021) Comparison of eeg signal characteristics between polysomnography and self applied somnography setup in a pediatric cohort. IEEE Access 9:110,916-110,926
Keil T, McBride D, Grimshaw K et al (2010) The multinational birth cohort of Europrevall: background, aims and methods. Allergy 65(4):482–490
Kemp B, Värri A, Rosa AC et al (1992) A simple format for exchange of digitized polygraphic recordings. Electroencephal Clin Neurophysiol 82(5):391–393
Korkalainen H, Aakko J, Nikkonen S et al (2019) Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. IEEE J Biomed Health Inform 24(7):2073–2081
Koutsourelakis I, Vagiakis E, Roussos C et al (2006) Obstructive sleep Apnoea and oral breathing in patients free of nasal obstruction. Europ Respir J 28(6):1222–1228
Le Nguyen T, Gsponer S, Ilie I et al (2019) Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations. Data Min Knowl Discov 33(4):1183–1222
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Lee SY, Guilleminault C, Chiu HY et al (2015) Mouth breathing, nasal disuse, and pediatric sleep-disordered breathing. Sleep Breath 19(4):1257–1264
Lemnaru C, Potolea R (2011) Imbalanced classification problems: systematic study, issues and best practices. In: international conference on enterprise information systems, Springer, pp 35–50
Leung K, Leckie C (2005) Unsupervised anomaly detection in network intrusion detection using clusters. Proc Twenty-Eighth Austral Conf Comput Sci 38:333–342
Li L, Yan J, Wang H et al (2020) Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder. IEEE Trans Neural Netw Learn Syst 32(3):1177–1191
Malhotra P, Vig L, Shroff G, et al (2015) Long short term memory networks for anomaly detection in time series. In: Proceedings, pp 89–94
Marcus CL (2001) Sleep-disordered breathing in children. Am J Respirat Crit Care Med 164(1):16–30
Markun LC, Sampat A (2020) Clinician-focused overview and developments in polysomnography. Curr Sleep Med Rep 6:309

Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorob 7:21
Oner MU, Cheng YC, Lee HK, et al (2020) Training machine learning models on patient level data segregation is crucial in practical clinical applications. medRxiv 2020–04
Park D, Hoshi Y, Kemp CC (2018) A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. IEEE Robot Autom Lett 3(3):1544–1551
Peralta M, Jannin P, Baxter JS (2021) Machine learning in deep brain stimulation: A systematic review. Artificial Intelligence in Medicine 122(102):198
Ratanamahatana CA, Keogh E (2005) Three myths about dynamic time warping data mining. In: proceedings of the 2005 SIAM international conference on data mining, SIAM, 506–510
Rewicki F, Denzler J, Niebling J (2023) Is it worth it? Comparing six deep and classical methods for unsupervised anomaly detection in time series. Appl Sci 13(3):1778
Ribeiro M, Lazzaretti AE, Lopes HS (2018) A study of deep convolutional auto-encoders for anomaly detection in videos. Patt Recogn Lett 105:13–22
Sabil A, Glos M, Günther A et al (2019) Comparison of apnea detection using oronasal thermal airflow sensor, nasal pressure transducer, respiratory inductance plethysmography and tracheal sound sensor. J Clin Sleep Med 15(2):285–292
Sano M, Sano S, Kato H et al (2018) Proposal for a screening questionnaire for detecting habitual mouth breathing, based on a mouth-breathing habit score. BMC Oral Health 18(1):1–13
Shwartz-Ziv R, Armon A (2022) Tabular data: Deep learning is not all you need. Information Fusion 81:84–90
Sigurdardóttir ST, Jonasson K, Clausen M et al (2021) Prevalence and early-life risk factors of school-age allergic multimorbidity: the europrevall-ifaam birth cohort. Allergy 76(9):2855–2865
Zhao B, Lu H, Chen S et al (2017) Convolutional neural networks for time series classification. J Syst Eng Electron 28(1):162–169

Authors and Affiliations

Luka Biedebach[1] · María Óskarsdóttir[1] · Erna Sif Arnardóttir[1] · Sigridur Sigurdardóttir[1] · Michael Valur Clausen[2] · Sigurveig Þ. Sigurdardóttir[2] · Marta Serwatko[2] · Anna Sigridur Islind[1]

✉ Luka Biedebach
lukab@ru.is

María Óskarsdóttir
mariaoskars@ru.is

Erna Sif Arnardóttir
ernasifa@ru.is

Sigridur Sigurdardóttir
sigridursig@ru.is

Michael Valur Clausen
mc@landspitali.is

Sigurveig Þ. Sigurdardóttir
veiga@landspitali.is

Marta Serwatko
martas@landspitali.is

Anna Sigridur Islind
islind@ru.is

[1] Reykjavik University, Menntavegur 1, 102 Reykjavik, Iceland

[2] Landspitali University Hospital, Hringbraut, 101 Reykjavik, Iceland

Appendix C

Publication III

Two Sides of the Same Pillow: Unfolding the Relationship between Objective and Subjective Sleep Quality with Unsupervised Learning

Complete Research Paper

Luka Biedebach
Reykjavik University
Menntavegur 1, 102 Reykjavik, Iceland
lukab@ru.is

María Óskarsdóttir
Reykjavik University
Menntavegur 1, 102 Reykjavik, Iceland
mariaoskars@ru.is

Erna Sif Arnardóttir
Reykjavik University
Menntavegur 1, 102 Reykjavik, Iceland
ernasifa@ru.is

Anna Sigridur Islind
Reykjavik University
Menntavegur 1, 102 Reykjavik, Iceland
islind@ru.is

Abstract

Advances in digital health allow us to take an active part in monitoring and improving our sleep quality. Both, objectively recorded and subjectively perceived sleep quality impacts our general health and well-being. This research shows how these two dimensions of sleep quality can be captured with smartwatches and digital symptom trackers. We contribute to the gap in the literature on how recorded values from wearables and user-generated content from mobile applications can elevate each other. Analysing the recorded and reported sleep quality in a longitudinal sleep study (n=45) shows differences in how participants perceive their sleep. We address this need for personalization, by creating clusters of participants with a similar perception of sleep using unsupervised machine learning. Analysing these clusters provides us with a more wholesome understanding of their sleep quality and raises awareness for the uniqueness of individuals in digital health.

Keywords: sleep quality, unsupervised machine learning, wearables, mobile application, clustering

Introduction

Our well-being depends on good sleep (Luyster et al., 2012), but the notion of sleeping well is still not fully understood (Buysse, 2014). Just like for any objectively and subjectively captured phenomenon, there are two sides of the same coin – or the same pillow – respectively for sleep quality. Bad sleep can have several negative effects on the body including daytime sleepiness, memory impairment (Orzeł-Gryglewska, 2010) and decreased neurobehavioral performance (Belenky et al., 2003). Chronically bad sleep increases the risk of inflammatory diseases (Irwin, 2015), cardiovascular diseases (Hoevenaar-Blom et al., 2011) and obesity (Beccuti & Pannain, 2011). There is no generally accepted definition of what sleep quality entails, or how to capture it. It can be described by characteristics of a person's sleep that can be measured or quantified by parameters such as the total sleep time, the sleep onset latency and sleep efficiency (Krystal & Edinger, 2008). The rise of digital health and the transition of digital health solutions from the clinic to the home

allows more people to track both their subjective and objective sleep quality. Objective sleep quality can be derived from tracking the body with sensors during sleep. The gold standard of sleep measurement is polysomnography, which continuously records the activity of the brain, the heart, the muscles and the respiratory system through multiple sensors during the night (Bruyneel et al., 2011). It is to date, seen as the the most reliable form of sleep measurement but it is not feasible for measuring sleep over an extended period of time due to its high effort. Wearable devices such as smartwatches measure the heart rate and movement, to give an estimation of a person's bed and rise times, the number of awakenings during the night and the duration of light sleep and deep sleep (Sadeh, 2011). Even though wearables measure sleep less reliably than a full polysomnography, their strength is to collect longitudinal data, the so-called 'free-living sleep' (Arnardottir et al., 2021), which is highly relevant in sleep research (Óskarsdóttir et al., 2022). However, sleep quality is a multidimensional construct that cannot be explained by sleep parameters only. The way we subjectively perceive our sleep, i.e. if we feel rested or fatigued, is essentially equally or more important than any measurement (Bin, 2016; Hoevenaar-Blom et al., 2011). Walsh et al. (2022) demonstrated the relevance of subjective sleep quality in the example of upper respiratory tract infections. Subjective sleep quality can for instance be measured as a Likert-style rating by the participant (Krystal & Edinger, 2008). On a more granular level, symptom trackers are used to split the rating into questions regarding the ease of falling asleep, the comfort during sleep or the feeling when waking up. This can be done digitally in the form of an entry in a digital symptom tracker application or in an analogue way via a manually filled-in sleep diary (Schmitz et al., 2022). Buysse (2014) introduced the highly relevant concept of sleep health, where objective and subjective sleep quality act combined as a wholesome indicator of sleep health. If we understand the underlying correlations, we can leverage the usage of wearables in combination with digital symptom trackers to objectively and subjectively monitor our health and well-being in general, and sleep in particular (Islind et al., 2022; Vallo Hult et al., 2022).

Digital health platforms and wearables allow patients to take part in the care process more actively and become co-creators of their own health data, which makes more individualized approaches for patient care possible (Topol, 2014). This leads to a shift towards precision medicine, which adjusts the medical treatment to the individual needs of the participants (National Research Council, 2011) and individualized, data-driven healthcare (Lillie et al., 2011). For sleep-disordered patients, the continuous form of health data collection enables personalized remote care by health professionals (Grisot et al., 2019). Another recent development in health information systems is self-monitoring, which enables individuals usually through wearables or symptom trackers to keep track of their symptoms and behaviors, review the self-recorded data and act accordingly. This may include adjusting their behaviour, applying treatment or seeking the help of a professional (Jiang & Cameron, 2020). Common motivations for self-tracking are curiosity, chronic disease or improving personal performance (Baumgart & Wiewiorra, 2016). Self-monitoring of sleep quality can help to increase the individuals awareness of sleep hygiene (Mairs & Mullan, 2015) and adjust their behaviour accordingly (Berryhill et al., 2020). However, there has been no research yet on how to combine objective and subjective measures for self-monitoring in sleep. For this reason, we aim to answer the following research question:

RQ: *How can the information from wearable devices and digital symptom trackers be combined to achieve more individualized sleep analysis?*

By answering this research question we contribute to the field of information systems, individualized healthcare and self-monitoring by showing the need for an individual-level analysis of sleep quality and proposing a method to address this need. We contribute to the field of digital health, especially regarding wearable devices and their health outcomes as well as data-driven approaches for digital subjective symptom trackers. Our main contributions are the analysis of subjective and objective sleep parameters and the proposed method of clustering participants based on their sleep perception for more individualized sleep analysis. The rest of this paper is organized as follows. In the next section, we review the related literature on the relationship between objective and subjective sleep quality and individualized healthcare. We then introduce the longitudinal sleep study in more detail and explain the methodological structure of this paper, including clustering and cluster analysis. We will present the results of the cluster analysis and based on the identified characteristics and correlations define *sleep quality types*. Finally, we will discuss the implications of our results in the context of digital health.

Related Work

Mobile digital health application and wearable devices have become a widespread and accepted way for tracking ones health (Farivar et al., 2020). The pandemic led to an increased speed in the adoption of digital health applications and devices which showed that the future of public health is becoming increasingly digital (Budd et al., 2020). Whether we model the health status of individuals through their continuous flow of data from digital health applications to personalize medical care or aggregate their data to generate new information about populations, there is potential for improving public health through digital health solutions (Kamel Boulos & Zhang, 2021). These digital health solutions could be mobile health applications (i.e., apps) which have been widely used for symptom tracking, implementing behavioural changes and remote care (Ghose et al., 2021). Moreover, wearable devices such as smartwatches have shown success in continuous long-term health monitoring (Dunn et al., 2018) and health education (Sultan, 2015). Additionally, there is existing research on combining the data from mobile apps and wearable devices. Sigurðardóttir et al. (2022) showed that leveraging objective data from smartwatches and subjective data from digital symptom trackers can enhance the clinical decision-making process for healthcare professionals for determining the ebb and flow in symptoms when treating patients with schizophrenia or bipolar disorders. Bremer et al. (2017) used machine learning methods on digital symptom trackers that produce diary data to predict the mood level of patients with depression. Their work showed how this kind of data can contribute to gaining insights on a subjective and therefore hard-to-measure condition with clinical relevance. It also showed how this subjective data can support personalized interventions. Another subjective phenomena discussed in digital health is subjective well-being. Hu et al. (2023) use a fitness health application to analyse the user's subjective well-being. Similar to subjective sleep quality it cannot be tied to an objective measurement, as it represents an individual's emotional responses and their general life satisfaction (Aboelmaged et al., 2021). The existing work in this area shows the potential of digital health for public health and possible solutions for handling the discrepancy between objective and subjective data. However, sleep - an essential prerequisite for health and well being - has not been approached with a combined analysis of objective and subjective longitudinal data as presented in this research. According to Bin (2016) there is a need to analyse effects of sleep quality on both physical and subjective health to understand the contribution of sleep as a whole to public health.

Previous research has attempted to analyse the relationship between objective and subjective sleep quality in general populations and found little correlation between the objective sleep parameters and subjective sleep quality (Baker et al., 1999). Zhang and Zhao (2007) raised the question whether subjective sleep quality relates to any objective sleep parameters, arguing that it is a combination represented by more than one parameter. The most important drivers of subjective sleep quality have been found to be the recorded total sleep time, awakenings, sleep efficiency (Åkerstedt et al., 2016) and reported sleep parameters (Goelema et al., 2019). Åkerstedt et al. (2016) showed that there are varying relationships between objective and subjective sleep quality in different age groups. Kaplan et al. (2017) approached the relationship between objective sleep parameters and subjective sleep quality with machine learning. They divided the population into age groups and then used a supervised classifier to predict the subjective sleep quality and analysed the feature importance. They showed that the correlations between objective sleep parameters and subjective sleep quality change throughout the age groups. On a more granular level, previous research did compare objective sleep parameters with the subjective reporting of sleep parameters. This gap, which is referred to as *sleep perception*, was previously considered as an effect of sleep disorders and medication as shown by Baker et al. (1999). Other researchers, such as Pinto et al. (2009) compared sleep perception of healthy and sleep disordered individuals and concluded that the way individuals perceive their sleep may be a relevant marker for the evolution of disorders and the effectiveness of treatment. Means et al. (2003) show that sleep perception varies amongst individuals and that distinctive subgroups with similar sleep perceptions can be identified with clustering. We aim to improve their approach by using an extended time period of sleep tracking through wearables instead of single nights from a traditional polysomnography recording. Based on these findings, we propose a method which analyses sleep in more homogeneous groups as proposed by Åkerstedt et al. (2016). In contrast to their work, we do not create those groups based on demographics such as age or gender but based on the gap between their objective and subjective parameters as proposed by Means et al. (2003).

Methods

The study was conducted within the research project Sleep Revolution (Arnardottir et al., 2022) at Reykjavik University in Iceland. The data for this research were collected by two different means simultaneously. Participants were asked to i) wear a Withings smartwatch (Issy-les-Moulineaux, France) and ii) fill out a digital sleep diary in the Sleep Revolution app (Reykjavik, Iceland), for 90 consecutive days. Collecting data from heterogeneous data sources is a challenging task. The data from the smartwatch and the digital sleep diary app were transformed into a homogeneous data format and stored in a digital platform as proposed by Sveinbjarnarson et al. (2023). The study is covered by ethical approval of the National Bioethics Committee of Iceland (21-070) and was approved by the Data Protection Agency of Iceland and includes a written consent by each participant. The participants were recruited through online and offline campaigns in the general population and were selected based on their age, BMI, gender and health status. The study aimed to include individuals with a wide range of age and body mass index and have a equal distribution of male and female participants as well as healthy and sleep disordered participants. Only participants who wore the smartwatch and filled out the digital sleep diary app for more than two weeks were selected for this analysis, which included 45 of 63 total participants. The participants wore the smartwatch on average for 75 days and filled out the digital sleep diary app on average for 40 days. This resulted in 2259 nights with information about both objective and subjective sleep quality in total. The population was gender balanced with 53.3% women. The average BMI was 28.2 and the average age was 48.9. There were 5 participants younger than 30, 32 participants between 30 and 60 and there were 8 participants older than 60. Most participants had a high educational level, as 64,4% of them have a university degree, 13,3% of them did vocational training or a technical degree and 22,2% have no degree in higher education. 3 of the 45 participants work in shift work or do night shifts. 80% of them were married or living with a partner and 93.3% were employed or studying. An overview of demographic information about the participants can be seen in Table 1.

Variable	Mean ±Standard Deviation
Age [years]	48.9 ±14.6
Body Mass Index [kg/m^2]	28.2 ±4.7

Table 1. Demographic Information

Objective Data: Smartwatch and Self-Applied Somnography

The smartwatch provides raw measurements, such as heart rate, oxygen saturation as well as aggregated information about the participant's sleep for each night. The aggregated data included the time the participant spent in light sleep and in deep sleep. The REM sleep, the sleep stage in which dreams are experienced, is not captured by this smartwatch even though it accounts for 20% to 25% of sleep (Carskadon, Dement, et al., 2005). The smartwatch also captures the time spent awake during the night the number of awakenings during the night. It measures the sleep onset latency, i.e the duration to fall asleep, and the duration to wake up. Being awake does not necessarily include standing up during the night, it also includes periods of laying in bed awake. Hence, awakenings can be several minutes or even hours, but in most cases only lasts for seconds and may not be remembered in the morning by the individual. The average sleep duration among all participants is 8.2 hours with a standard deviation of 1 hour. Other sleep parameters have been calculated from the available data. This includes sleep efficiency, which is the time asleep during the night divided by the total time spent in bed. Sleep variability was calculated, by taking the standard deviation of the sleep duration of the previous 3 nights. During the night, the heart rate was tracked and aggregated as minimum, maximum and average heart rate. The smartwatch tracked the participants not only during the night but also tracked their activity during the day. It captured the participant's number of steps, distance and elevation. The average distance among all participants is 3.1 km per day with a standard deviation of 1.8. We transformed all time-related features into seconds and brought all features to a uniform scale using the scikit-learn StandardScaler. An overview of all smartwatch variables can be found in Table 2.

Additionally, each participant participated in a self-applied somnography. Somnography is a simplified version of a polysomnography, which is designed for measuring sleep outside of the hospital but with a

similar reliability to the traditional measurement (Kainulainen et al., 2021). The measurement included an electroencephalogram (EEG), electrooculogram (EOG), electrocardiogram (ECG), RIP belts for the thorax and abdomen, a finger probe pulse oximeter, a microphone, a nasal cannula, electrodermal activity (EDA) and accelerometry measuring the movement and body position. The somnography provides a more reliable measurement of the participants than the smartwatch but was, due to its high measurement effort, only performed on up to three consecutive nights whereas the smartwatch was worn for the 90-day period. There are 3 participants with no somnography recorded nights, 3 participants with one recorded night, 11 participants with two recorded nights and 28 participants with three recorded nights. All recordings have been manually scored by sleep technologists according to the rules of the American Academy of Sleep Medicine (Berry et al., 2018).

Subjective Data: Digital Sleep Diary from a Mobile Application

The subjective sleep quality was captured with a digital sleep diary in the form of the custom-made app developed by Sleep Revolution. The sleep diary was co-designed and developed, according to the digital sleep diary standards proposed by Schmitz et al. (2022). The design aimed to increase compliance with the digital sleep diary and avoid memory bias. The sleep diary was filled out by the participant two times per day. In the morning diary, the participants described their nocturnal sleep. Additionally, they reported sleep parameters such as the total sleep time, the sleep onset latency, the number of awakenings and the awake time during the night. They were reminded to fill out the digital sleep diary through nudging, delivered by push notifications. In the evening diary, the participants filled in information about their day, in particular about the factors which can impact sleep. This included their stress level, daytime sleepiness, naps, the number of caffeine and alcohol units they consumed and the duration of exercise. This data was used to measure the participants' subjective sleep quality. The average sleep quality rating among all participants was 3.3 out of 5 with a standard deviation of 0.5.

Source	Variables
Smartwatch	Sleep hour, wake-up hour, variability, efficiency, regularity, light sleep, deep sleep, duration to wake up, heart rate (avg, min, max), steps, distance, elevation
Sleep Diary	Exercise duration, work day, stress level, nap count, nap duration, drug use, alcohol count, caffeine count
Both	Sleep duration, awake time, awakenings, sleep onset latency

Table 2. Variables from the Smartwatch and Sleep Diary

Data Analysis

The methods are structured in three parts as can be seen in Figure 1. First, the agreement between recorded and reported sleep parameters was analysed. We calculated the agreement of the objective sleep parameters recorded with the smartwatch and the subjective sleep parameters reported in the sleep diary for each participant. The results from this step described how the individual participants perceived their own sleep. Based on this information, the second step was to cluster the participants into groups with similar sleep perceptions, using the unsupervised learning algorithm K-Means. Lastly, we analysed these clusters of participants with different sleep perceptions. We identified characteristics of the clusters using an analysis of variance. This allowed us to define sleep quality types. Finally, we analysed the correlations between factors that can impact sleep quality in order to understand which areas of improvement are relevant for which sleep quality type.

Comparison of Reported and Recorded Sleep Parameters

There are multiple sleep parameters that were captured both objectively and subjectively. They were recorded by the smartwatch and reported in the digital sleep diary app. This included sleep duration, the sleep onset

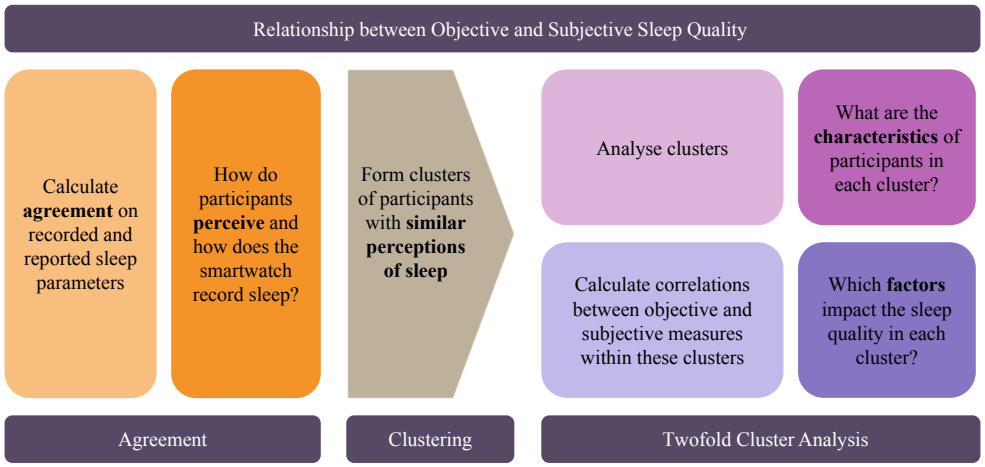


Figure 1. Methodological Structure

latency, the number of awakenings and the awake time after sleep onset as can be seen in Table 2. We directly calculated the agreement of the reported and recorded parameters by taking the absolute difference for each of the four overlapping sleep parameters for each night. This comparison showed us the gap between objective recording and subjective reporting of each participant. It remains unknown whether this gap arises from unreliable recording or reporting, but it indicates how the recorded sleep is connected to the individual perception of sleep of the individual participants. We took neither the smartwatch recording, nor the participants' reporting as ground truth, instead this comparison gave us information about how related or unrelated these two measures generally are. An overview of the availability of both recorded and reported sleep parameters of all participants can be seen in Figure 2. It shows that most participants filled out the sleep diary regularly, with their rating represented as colorful dots. Moreover, it shows that the rating behavior of participants varies, as some participants gave more extreme ratings such as 1 and 5, while others mainly rated medium sleep quality between 2-4.

Additionally, we reviewed the general reliability of the recorded and reported sleep parameters with somnography recordings, which is seen as a more accurate tool for sleep measurement, although not longitudinal. The somnography sleep parameters were manually scored by sleep technologists. The participants had up to three nights of somnography recording, smartwatch recording and digital sleep diary reporting simultaneously. This review gave us important information about the reliability of the recorded and reported sleep parameters, but ultimately the aim of this proposed method is to rely only on recorded data from wearable devices in combination with reported data from a digital symptom tracker.

Clustering

Using unsupervised machine learning, we aimed to identify clusters among participants to find out what impacts the perception of their sleep quality. Clustering allows a more individual-level analysis of the data by creating clusters of participants in a way that participants within a group are more similar to each other than to participants in the other clusters. This strategy can be used to e.g. characterize clinical phenotypes (An et al., 2020). We use the partitional clustering algorithm *K-Means* to identify similar groups of participants. We chose K-means, because it is works well for balanced cluster sizes and small cluster numbers. It furthermore has a low computation time and is easily understandable. The K-means algorithm assigns each

observation to the cluster with the closest mean over multiple iterations (MacQueen, 1967). The value of *K* defines the number of clusters. The optimal value for *K* is chosen by the steepest descent in an elbow plot, which shows the cluster impurity by the number of clusters. We performed the clustering on the calculated agreement between the recorded and reported sleep parameters and the subjective sleep quality.

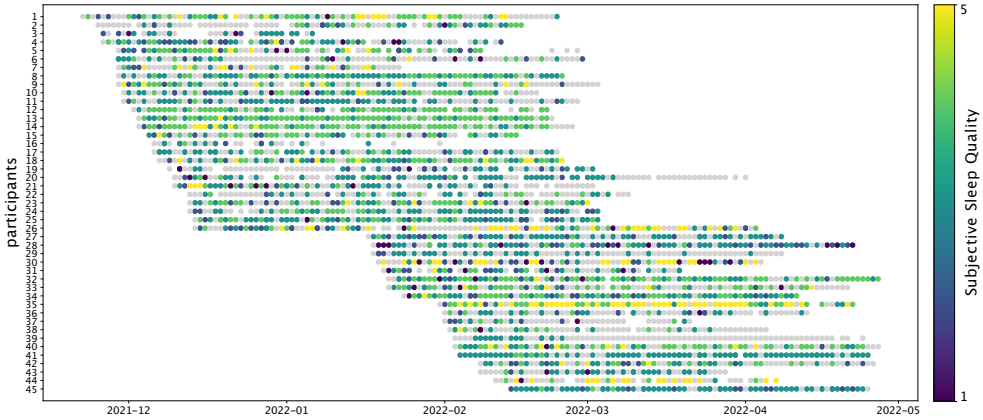


Figure 2. Study Coherence by the Participants over Time (every dot represents a night in which the smartwatch was worn, colored dots describe the subjective sleep quality from the sleep app and grey dots represent recorded nights without a sleep quality rating)

The cluster analysis was twofold, i) we identified common characteristics of the participants within each cluster and ii) we calculated the correlations between the variables and the subjective sleep quality within each cluster. This allowed us to define sleep quality types and improve the understanding of the different factors impacting sleep quality within the clusters. We first analysed the clusters by comparing the median and standard deviation of each cluster. The aim of this analysis was to identify common characteristics of participants within the clusters. Additionally, an analysis of variance (ANOVA) showed, whether the difference between the groups was significant. We performed the ANOVA for each feature to determine the F statistic and p-value. As a second step, we calculated Spearman's rank correlation coefficient between possible impacting factors and the subjective sleep quality over the whole study duration within each cluster. This method resulted in a *rho* value between -1 and 1 for each variable, representing the positive or negative correlation with the subjective sleep quality. Additionally, it resulted in a p-value for each correlation, representing the significance of the correlation. In this analysis, we only considered correlations with a p-value lower than 0.05 as significant.

Results

In the following section, we review the results of our analysis according to the different steps of our method. We first show the gap between the reported and recorded sleep parameters, then introduce the four clusters that were identified in the clustering and then show characteristics of the clusters and review the individual correlations to subjective sleep quality in each cluster.

Sleep Parameter Agreement

In order to understand the relationship between objective and subjective sleep quality, we first compared sleep parameters that were simultaneously captured with the smartwatch, sleep diary and somnography

recording for up to three consecutive nights. This comparison showed that there is a varying agreement of objective and subjective sleep quality among the participants. Figure 3 shows the sleep duration of the smartwatch, digital sleep diary app and somnography in 15 exemplary participants. We can see that each participant has a different degree of agreement between the three reported sleep parameters. Considering somnography as the most reliable measurement, we can see that for 30% of the participants the reported sleep duration is more accurate than the recorded sleep duration, while for 70% the recorded sleep duration is more accurate.

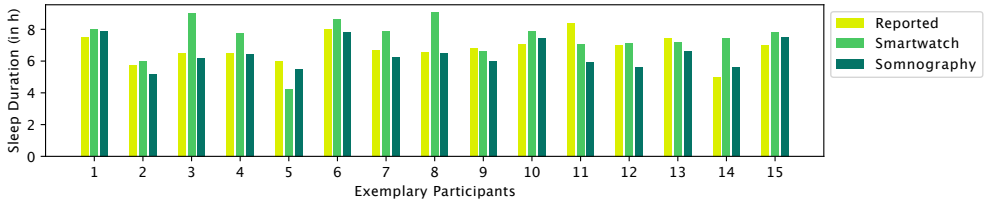


Figure 3. Comparison of Reported Sleep Duration to the Smartwatch and Somnography

Figure 4 shows the difference between the reported and recorded sleep duration. All data points above the vertical line are nights where the participant reported a higher sleep duration than the smartwatch. All nights under the vertical line have a lower reported sleep duration than recorded sleep duration. The color of the data points reflects the subjective sleep quality assigned to the night by the participant. Here, we can see that the subjective sleep quality is usually lower when the reported sleep duration is low regardless of the recorded sleep duration. The reported sleep duration by the participants is on average 57 minutes lower than the recorded sleep duration by the smartwatch. The participant with the highest difference between recorded and reported sleep duration reports on average two hours less than captured by the smartwatch.

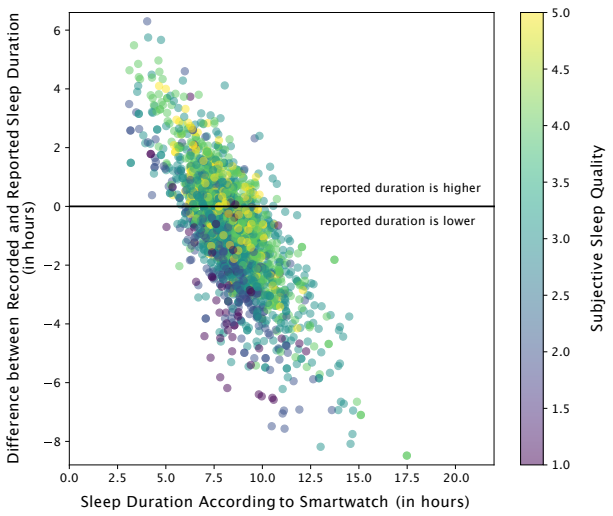


Figure 4. Difference Between Recorded and Reported Sleep Duration

The average number of recorded awakenings during the night is 1.6 with a standard deviation of 1.7. The reported number of awakenings is on average 0.26 lower than the recorded value. An increased number of recorded awakenings leads to a more extreme difference between the recorded and the reported value. Comparing the number of awakenings between the smartwatch and the somnography shows that the recorded value by the smartwatch is usually too high. The sleep onset latency has the proportionally lowest agreement. A comparison of recorded and reported values show that the smartwatch usually records a lower sleep onset latency than the participants report. The average sleep onset latency recorded by the watch is 3 minutes, while the average reported time by the participant is 18 minutes.

Sleep Quality Types

We found the optimal number of clusters by creating an elbow plot, that shows the intra-cluster similarity by the number of clusters. The steepest drop of the curve can be observed at four clusters, which is why we chose K=4. The first cluster has 11, the second cluster 8, the third cluster 11 and the fourth cluster has 15 participants. We use the identified clusters to define sleep quality types based on their agreement, characteristics and correlations. In the 1970s, Horne and Östberg designed a questionnaire to assess an individual's sleep type according to their circadian rhythm (Horne & Östberg, 1976). Based on this research, sleep chronotypes have been developed, which provide personal guidance on the optimal bed and rise times, as well as the optimal time of productivity during the day (Roenneberg et al., 2003). In contrast to the chronotypes, the sleep quality types proposed in this research aim to show influencing factors for the personal perception of sleep quality. The defined sleep quality types can be seen in Figure 5.

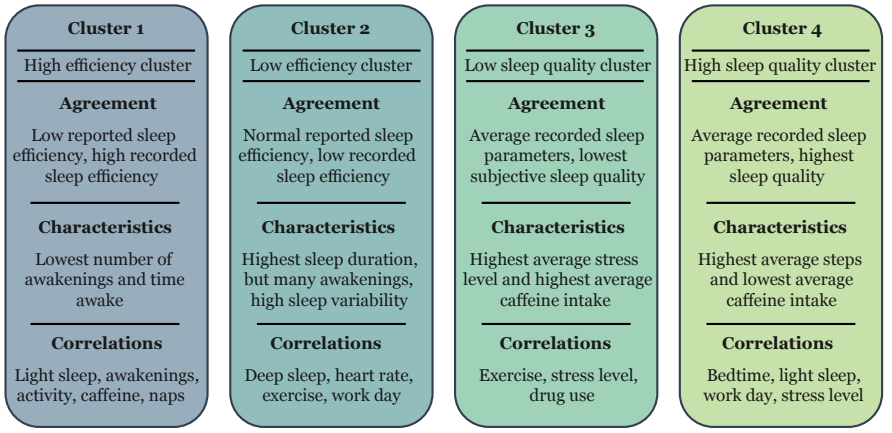


Figure 5. The Four Identified Sleep Quality Types

Identified Characteristics within the Clusters

We use box plots to visually identify characteristics of the different clusters. We show the distributions in each cluster and additionally compare them to the population mean. There is an even distribution of age, body mass index (BMI) and gender throughout the clusters. The main differences we identified between the clusters were found in their relationship between the recorded sleep parameters and sleep quality in general and the relationship between reported and recorded sleep efficiency. Since sleep efficiency describes the sleep duration relative to the total time in bed, a low sleep onset latency and a low duration of time awake during the night contribute to a high sleep efficiency regardless of the total sleep duration. Therefore, sleep efficiency is a combination of all 4 sleep parameters. The sleep quality types in Figure 5 show how the main difference between cluster 1 and 2 are the opposing gaps between reported and recorded sleep efficiency.

Figure 6 shows how the clusters significantly differ in their median subjective sleep quality. Cluster 4 reported on average a higher sleep quality than all other clusters, even though its median sleep duration is close to the population median of 8.2 h. What distinguishes cluster 4 is the relatively high reported sleep duration in comparison to the recorded sleep duration. Cluster 1 has an almost identical distribution of recorded and reported sleep duration but does not show an increased subjective sleep quality.

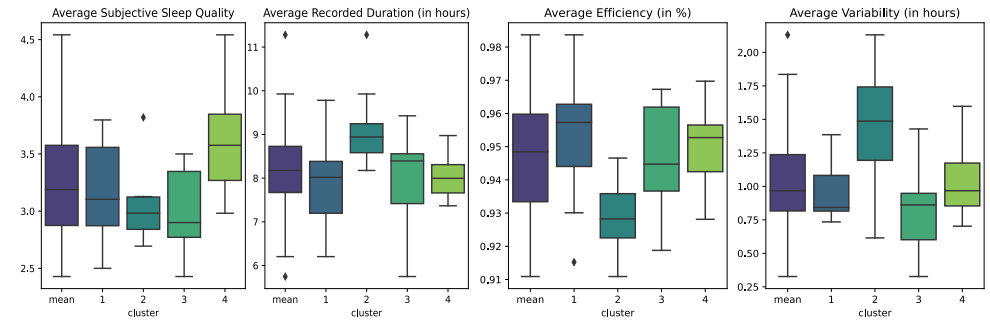


Figure 6. Distribution of Reported Sleep Quality, Recorded Sleep Duration, Recorded Sleep Efficiency and Recorded Sleep Variability among the Clusters

Cluster 1 and two show differences both in the reported and recorded sleep efficiency. Cluster 1 shows a high recorded sleep efficiency, which does not match the high number of reported awakenings and awake time after sleep onset. Contrarily, has cluster 2 the lowest median recorded sleep efficiency, even though the reported sleep parameters indicate a normal sleep efficiency. This cluster has the highest median sleep duration of 9 hours but still has a below-average subjective sleep quality. We conclude that the subjective sleep quality of cluster 2 may be mainly impacted by sleep efficiency.

Avoiding stress, being active, avoiding caffeine after midday and avoiding alcohol can be beneficial for your sleep. Cluster 4, the high subjective sleep quality cluster, shows the highest median recorded activity during the day and the lowest median daily caffeine intake. Cluster 3, the cluster with the lowest sleep median subjective sleep quality, has the highest median stress level and highest median daily caffeine intake. Figure 6 shows that the second cluster has the highest variability of all clusters with a median of 1.5 hours. The participant's medical background gives us a further understanding of the identified characteristics among the clusters. Figure 7 shows the prevalence of moderate and severe insomnia within each cluster according to the insomnia severity index (Morin et al., 2011). The cluster with low sleep efficiency and high sleep variability has a high percentage of insomnia patients, while the cluster with the highest sleep quality has the lowest percentage of insomnia patients.

Identified Correlations within the Clusters

Figure 8 shows the correlations of each variable to the subjective sleep quality for the participants in each cluster. Warm colors indicate a positive correlation and cold colors indicate a negative correlation. Black indicates no correlation between the variable and the subjective sleep quality. We exclude all non-significant correlations with a p-value higher than 0.05 from the analysis and set them to zero. All clusters show the strongest correlation between subjective sleep quality and the reported sleep parameters. These variables have been reported in the same way and at the same time as the subjective sleep quality, which may explain the high correlation. Figure 8 shows that all clusters have different correlations between the given variables and the subjective sleep quality. Cluster 1 shows positive correlations with both sleep duration and wake-up time. This means, that a later bedtime in the morning and a longer total sleep duration are associated with higher subjective sleep quality. Cluster 1 is the only cluster that shows a negative correlation between

exercise and subjective sleep quality. Surprisingly, we can see a positive correlation with caffeine intake. Cluster 2 has a negative correlation between subjective sleep quality and heart rate. A high average heart rate during the night is associated with lower sleep quality in this cluster. Additionally, cluster 2 shows the highest positive correlation between exercise duration and subjective sleep quality. Therefore, more exercise is associated with higher subjective sleep quality in this cluster.

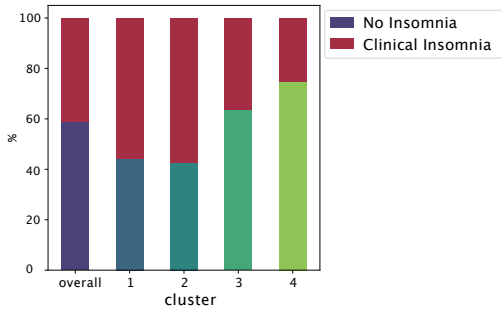


Figure 7. Insomnia Severity among the Clusters

Cluster 3 has almost no correlation with any variable from the smartwatch. The most relevant correlations to subjective sleep quality in this cluster are the four estimated sleep measures. This cluster shows higher correlations of this group of variables than all other clusters. This could indicate that a smartwatch is not a suitable measurement device for this cluster. Cluster 4, the high sleep quality cluster, has a positive correlation between bedtime and subjective sleep quality and has, with 12 PM, the latest median bedtime hour. Furthermore, we can see that a higher percentage of deep sleep correlates with higher subjective sleep quality. Similar to cluster 2, work days and stress are correlated with low subjective sleep quality.

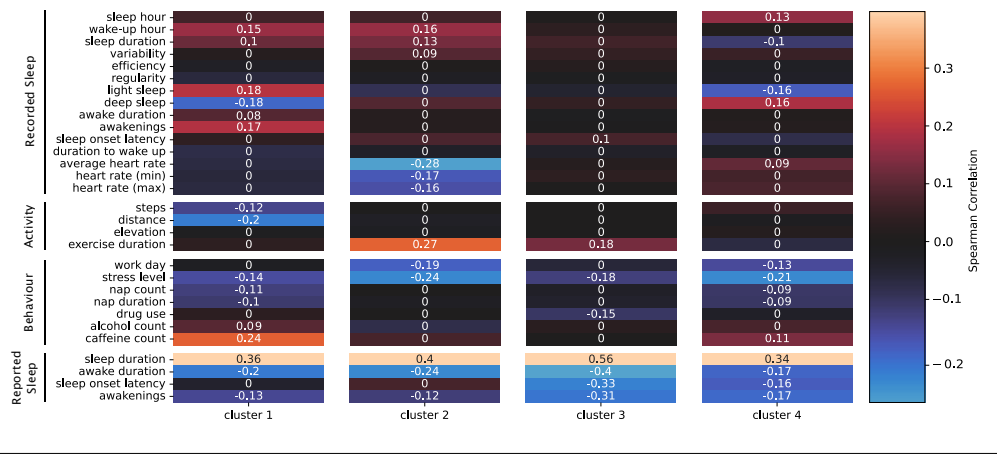


Figure 8. Correlations Between the Objective Attributes and the Subjective Sleep Quality within Each Cluster (non-significant correlations are displayed as zero)

If we take this personalized analysis one step further and look at the data of each participant individually even more extreme correlations become visible. Figure 9 shows a correlation matrix with each row representing one participant and each column representing one variable. The top row represents the full study population, in which only slight correlations are visible. The participants show individual correlations. We can see that most variables both a positive correlation with subjective sleep quality in some participants and a negative correlation in others. It further shows, that some participants have multiple strong correlations with subjective sleep quality while other participants show none. This perspective supports our results on the individuality of the relationship between objective and subjective sleep quality.

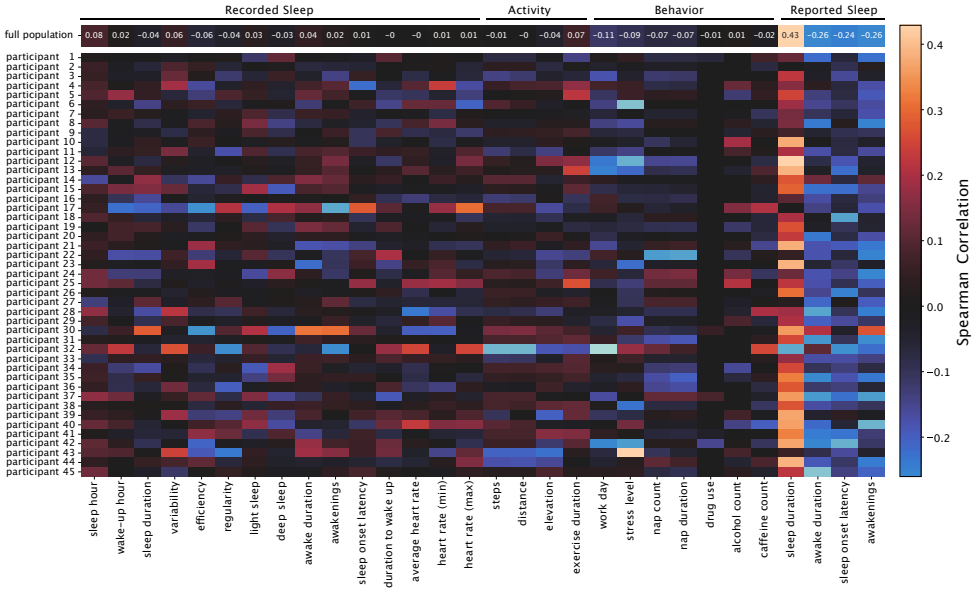


Figure 9. Correlations to Subjective Sleep Quality on the Population vs Individual Level

Discussion

The results presented in this paper showed that there are gaps between the recorded and reported sleep parameters. Moreover, our findings illustrate that some individuals have higher gaps between their reported and recorded sleep duration compared to others. This result could either be explained by measurement difficulties attributed to the smartwatch or to a form of insomnia, as affected individuals have been shown to have a strong discrepancy between objective and subjective sleep duration (Rezaie et al., 2018). The results furthermore show that there are big gaps between the recorded and reported sleep onset latency. When analysing that, we need to consider, that falling asleep is a continuous process where the participant is shifting between levels of consciousness, which makes it challenging both for the participant to give an estimation of the time and for the smartwatch to determine the sleep onset latency. The agreement of the number of awakenings is low across most participants. For all participants, the smartwatch records a higher number of awakenings than reported by the participants. Most awakenings during the night are short and happen during low consciousness, which makes it difficult to remember them. However, a smartwatch has a limited ability to capture awakenings as well (Gruwez et al., 2019). Overall, comparing the sleep parameters showed that both the reported sleep parameters from digital sleep diary app and the smartwatch deviate from the

measurements captured by the somnography. This confirms our assumption that neither the recorded nor the reported sleep parameters can be treated as a ground truth. However, for longitudinal sleep measurements, the combination of both, is a viable option.

We showed that the information inherent in these gaps can be used to create clusters of participants with similar sleep perception. The participants within these clusters showed commonalities regarding their median sleep quality, sleep duration, sleep variability and sleep efficiency. This result showed, that even though two clusters show a similar median sleep duration there are still differences in the subjective sleep quality. This is in line with the assumption of Bin (2016), who maintains that more dimensions of sleep quality than sleep duration need to be considered. Analysing these clusters of participants with similar sleep perceptions made correlations between exercise and subjective sleep quality visible. We observed that exercise only has a positive correlation in two clusters. Generally, exercise is beneficial for sleep, except when done shortly before sleep (American Academy of Sleep Medicine, 2005). Even though Hynynen et al. (2010) observe an increased nocturnal heart rate after moderate or heavy exercise during the day, Myllymäki et al. (2012) do not observe a negative impact on subjective sleep quality. As we only observe these correlations in one cluster, this effect might be overlooked when studying the general population. The regularity of sleep does not have a positive or negative correlation in any cluster, even though a positive correlation would be expected. Keeping to a regular sleep duration is beneficial, as high sleep variability has a negative effect on subjective well-being (Lemola et al., 2013) and increases the risk of weight gain (Kobayashi et al., 2013). In two clusters, the caffeine intake shows a positive correlation. This outlines a counterintuitive result, as caffeine is typically considered to have a negative effect on sleep. However, O'Callaghan et al. (2018) describe the complex cyclic relationship between caffeine consumption and sleep deprivation. This correlation might therefore arise from interactions, which are not included in this model, e.g., we do not include the time of consumption or the effects of caffeine withdrawal from *caffeine natives* (O'Callaghan et al., 2018). Finally, when looking at the correlations of each individual participant, the results show an even stronger differences across individuals. This shows that sleep quality is highly individual and based on that, we would like to emphasize the relevance of personalized sleep analysis for future research endeavours.

This paper contributes to the field of information systems in several ways. Firstly, by outlining a method for personalized digital health monitoring utilizing a combination of recorded data from wearable devices and reported data from digital self-tracking applications for the general population. Secondly, we show that the user-generated health data from smartwatches and digital symptom trackers does not necessarily reflect the sleep parameters derived from the somnography, but still gives us valuable information about the participant's sleep health. Thirdly, our paper shows that by combining both objective and subjective measurements and learning about the participant's sleep perception through the agreement between the two, we can develop personalized sleep interventions and go from short term monitoring to reliable, dimensional longitudinal data collection. Improving sleep monitoring with the combination of objective data from wearable devices and subjective data from self-tracking apps allows individuals to actively partake in their own health through technological interfaces, which Petrakaki (2017) refers to as *technological self-care*. Additionally, our paper makes contributions to the field of sleep research by comparing subjective and objective sleep quality in a longitudinal study. To the best of our knowledge, there is no comparable study that captures both the recorded and reported sleep parameters over an extended period of time. It showed that the sleep duration recorded by the smartwatch is on average higher than the reported sleep duration. Our results go in line with the results by Rupp and Balkin (2011) when comparing different wearables to a polysomnography. However, we additionally showed that participants tend to report lower sleep parameters than the smartwatch when they experience low sleep quality. As stated earlier in this paper, the potential of digital symptom trackers is seemingly large. Our research confirms that including information from the digital sleep diary app to the smartwatch measurements enhances sleep analysis. More specifically, we illustrate an added value in including both smartwatch and user-generated health data, since both objective and subjective sleep quality have clinical relevance. Finally, we show, that the interaction of these two dimensions of sleep quality creates an additional value in itself. The agreement between objective and subjective sleep quality contains information about the individual's sleep perception; the combination of the two, outlines two sides of the same pillow.

Practical Implications

The practical implications of our research are two-fold by showing: i) the value of combining objective data and user-generated health data and ii) the need for personalized sleep analysis to cater to the fact that there are individual differences that are important to consider. Current methods of sleep tracking usually either favor subjective or objective sleep data. As a contrast to that, we proposed that both are combined in order to gain an in-depth view of an individual's sleep. This method could be transferred to other areas of information systems and data-driven healthcare for contexts such as chronic disease management, where various types of health data exist, but a coherent integration framework between them is needed (Bardhan et al., 2020). In addition to that, we have illustrated interesting correlations between objective sleep parameters and subjective sleep quality, which may be relevant in clinical practice. We showed that the relationship between objective and subjective sleep quality varies between participants and because of that, we proposed to analyse sleep on in a more personalized manner, through our individual based method. Moreover, our findings show that general assumptions about sleep quality may not apply to all individuals and that issue could be addressed by analysing clusters of individuals with similar sleep perceptions.

Limitations and Future Research

One limitation of this study is the low reliability of the smartwatch, as they can only give estimations of the actual sleep, due to the placement of the sensor on the wrist, and the infancy of the technology to date. Therefore, measurement errors might prevent us from estimating the true sleep parameters. They have shown to be suitable to assess bed and rise times but show a low agreement with awakenings during the night or sleep efficiency (Sadeh, 2011). Kang et al. (2017) showed that smartwatches are less reliable for individuals with sleep insomnia. Furthermore, this work is based on data that is rarely collected over a long period of time. We hope that similar studies can be conducted in the future, as one limitation of the current research is the small number of participants in the study. It limits our ability to observe patterns across the study population. In future research, applying the clustering method to a larger study population may result in more significant characteristics of the clusters. Moreover, this method could be extended by taking the temporal dimension into account and learning from the changes in the participants' sleep over time. Similar to Liang et al. (2016), who proposed a method for calculating individual markers for good sleep quality, future research could do a dynamic analysis of sleep quality based on the previous nights.

Conclusion

This research showed, that the relationship between objective and subjective sleep quality is different for every individual. Comparing the sleep parameters resulting from the smartwatch and the digital sleep diary to the somnography showed that the reliability of both the recorded and reported parameters varied among the participants. We used this difference between reported and recorded sleep parameters to assess the participant's perception of sleep. Clustering participants with similar sleep perceptions allowed us to perform a more personalized analyse of their sleep quality. Based on the identified commonalities in sleep, activity and daytime behavior of these clusters we defined four sleep quality types. All clusters show different correlations between objective and subjective sleep parameters and subjective sleep quality. Hence, objective and subjective sleep quality, are two sides of the same *pillow*, and looking at both sides is vital for assessing sleep quality. Based on that, we propose to combine the data from both smartwatches and digital symptom trackers to outline an individual's perception of sleep. This allows for a more personalized sleep monitoring and ultimately more individual, data-driven patient care in the future.

Acknowledgements

We want to thank everyone who was involved in the set-up of the study. A special thanks to Dimitri Ferretti, Elena Richert, Lisa Schmitz, Bjarki Freyr Sveinbjarnarson, Emil Harðarson, Kolfinna Þórisdóttir, Jacopo Piccini and Katrín Ýr Friðgeirsdóttir. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965417. The corresponding author is Luka Biedebach and the senior author of this paper is Anna Sigríður Islind.

References

Aboelmaged, M., Hashem, G., & Mouakket, S. (2021). Predicting subjective well-being among mHealth users: a readiness–value model. *International Journal of Information Management* (56), 102247.

Åkerstedt, T., Schwarz, J., Gruber, G., Lindberg, E., & Theorell-Haglöw, J. (2016). The relation between polysomnography and subjective sleep and its dependence on age–poor sleep may become good sleep. *Journal of Sleep Research* (25:5), 565–570.

American Academy of Sleep Medicine (2005). International classification of sleep disorders. *Diagnostic and coding manual*, 51–55.

An, H.-J., Baek, S.-H., Kim, S.-W., Kim, S.-J., & Park, Y.-G. (2020). Clustering-based characterization of clinical phenotypes in obstructive sleep apnoea using severity, obesity, and craniofacial pattern. *European journal of orthodontics* (42:1), 93–100.

Arnardottir, E. S., Islind, A. S., Óskarsdóttir, M., Ólafsdottir, K. A., August, E., Jónasdóttir, L., Hrubos-Strøm, H., Saavedra, J. M., Grote, L., Hedner, J., et al. (2022). The Sleep Revolution project: the concept and objectives. *Journal of sleep research* (31:4), e13630.

Arnardottir, E. S., Islind, A. S., & Óskarsdóttir, M. (2021). The future of sleep measurements: a review and perspective. *Sleep medicine clinics* (16:3), 447–464.

Baker, F. C., Maloney, S., & Driver, H. S. (1999). A comparison of subjective estimates of sleep with objective polysomnographic data in healthy men and women. *Journal of psychosomatic research* (47:4), 335–341.

Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. *MIS Quarterly* (44:1), 185–200.

Baumgart, R. & Wiewiorra, L. (2016). The Role of Self-Control in Self-Tracking. *Proceedings of the International Conference on Information Systems (ICIS), Dublin, Ireland, December 11-14, 2016*.

Beccuti, G. & Pannain, S. (2011). Sleep and obesity. *Current opinion in clinical nutrition and metabolic care* (14:4), 402.

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., & Balkin, T. J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of Sleep Research* (12:1), 1–12.

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. M., & Vaughn, B. V. (2018). *AASM Manual for the Scoring of Sleep and Associated Events*. Tech. rep. Amer. Acad. Sleep Med., Darien, IL, USA: American Academy of Sleep Medicine, Version 2.5.

Berryhill, S., Christopher, J., Dean, A., Provencio-Dean, N., Patel, S. I., Estep, L., Combs, D., Mashaqi, S., Gerald, L. B., et al. (2020). Effect of wearables on sleep in healthy individuals: a randomized crossover trial and validation study. *Journal of Clinical Sleep Medicine* (16:5), 775–783.

Bin, Y. S. (2016). Is sleep quality more important than sleep duration for public health? *Sleep* (39:9), 1629–1630.

Bremer, V., Becker, D., Funk, B., & Lehr, D. (2017). Predicting the individual mood level based on diary data. *Proceedings of the 25th European Conference on Information Systems (ECIS), Guimarães, Portugal, June 5-10*, 1161–1177.

Bruyneel, M., Sanida, C., Art, G., Libert, W., Cuvelier, L., Paesmans, M., Sergysels, R., & Ninane, V. (2011). Sleep efficiency during sleep studies: results of a prospective study comparing home-based and in-hospital polysomnography. *Journal of sleep research* (20:1pt2), 201–206.

Budd, J., Miller, B. S., Manning, E. M., Lampos, V., Zhuang, M., Edelstein, M., Rees, G., Emery, V. C., Stevens, M. M., Keegan, N., et al. (2020). Digital technologies in the public-health response to COVID-19. *Nature medicine* (26:8), 1183–1192.

Buysse, D. J. (2014). Sleep health: can we define it? Does it matter? *Sleep* (37:1), 9–17.

Carskadon, M. A., Dement, W. C., et al. (2005). Normal human sleep: an overview. *Principles and practice of sleep medicine* (4:1), 13–23.

Dunn, J., Runge, R., & Snyder, M. (2018). Wearables and the medical revolution. *Personalized medicine* (15:5), 429–448.

Farivar, S., Abouzahra, M., & Ghasemaghaei, M. (2020). Wearable device adoption among older adults: A mixed-methods study. *International Journal of Information Management* (55), 102209.

Ghose, A., Guo, X., Li, B., & Dang, Y. (2021). Empowering Patients Using Smart Mobile Health Platforms: Evidence from a Randomized Field Experiment. *Forthcoming at MIS Quarterly, NYU Stern School of Business Forthcoming*.

Goelema, M., Regis, M., Haakma, R., Van Den Heuvel, E., Markopoulos, P., & Overeem, S. (2019). Determinants of perceived sleep quality in normal sleepers. *Behavioral sleep medicine*, 388–397.

Grisot, M., Moltubakk Kempton, A., Hagen, L., & Aanestad, M. (2019). Data-work for personalized care: Examining nurses' practices in remote monitoring of chronic patients. *Health informatics journal* (25:3), 608–616.

Gruwez, A., Bruyneel, A.-V., & Bruyneel, M. (2019). The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. *PLoS One* (14:1), e0210569.

Hoevenaar-Blom, M. P., Spijkerman, A. M., Kromhout, D., van den Berg, J. F., & Verschuren, W. (2011). Sleep duration and sleep quality in relation to 12-year cardiovascular disease incidence: the MORGEN study. *Sleep* (34:11), 1487–1492.

Horne, J. A. & Östberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International journal of chronobiology*.

Hu, J., He, W., Zhang, J., & Song, J. (2023). Examining the impacts of fitness app features on user well-being. *Information & Management* (60:5), 103796.

Hynynen, E., Vesterinen, V., Rusko, H., & Nummela, A. (2010). Effects of moderate and heavy endurance exercise on nocturnal HRV. *International Journal of Sports Medicine* (31:06), 428–432.

Irwin, M. R. (2015). Why sleep is important for health: a psychoneuroimmunology perspective. *Annual Review of Psychology* (66), 143–172.

Islind, A. S., Hult, H. V., Rydenman, K., & Wekell, P. (2022). Co-creating a Digital Symptom Tracker: An App as a Boundary Object in the Context of Pediatric Care. *International Working Conference on Transfer and Diffusion of IT,* Springer, 79–93.

Jiang, J. & Cameron, A.-F. (2020). IT-Enabled Self-Monitoring for Chronic Disease Self-Management: An Interdisciplinary Review. *MIS quarterly* (44:1).

Kainulainen, S., Korkalainen, H., Sigurðardóttir, S., Myllymaa, S., Serwatko, M., Sigurðardóttir, S. Þ., Clausen, M., Leppänen, T., & Arnardóttir, E. S. (2021). Comparison of EEG Signal Characteristics Between Polysomnography and Self Applied Somnography Setup in a Pediatric Cohort. *IEEE Access* (9), 110916–110926.

Kamel Boulos, M. N. & Zhang, P. (2021). Digital twins: from personalised medicine to precision public health. *Journal of personalized medicine* (11:8), 745.

Kang, S.-G., Kang, J. M., Ko, K.-P., Park, S.-C., Mariani, S., & Weng, J. (2017). Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *Journal of Psychosomatic Research* (97), 38–44.

Kaplan, K. A., Hardas, P. P., Redline, S., Zeitzer, J. M., Group, S. H. H. S. R., et al. (2017). Correlates of sleep quality in midlife and beyond: a machine learning analysis. *Sleep medicine* (34), 162–167.

Kobayashi, D., Takahashi, O., Shimbo, T., Okubo, T., Arioka, H., & Fukui, T. (2013). High sleep duration variability is an independent risk factor for weight gain. *Sleep and Breathing* (17:1), 167–172.

Krystal, A. D. & Edinger, J. D. (2008). Measuring sleep quality. *Sleep medicine* (9), S10–S17.

Lemola, S., Ledermann, T., & Friedman, E. M. (2013). Variability of sleep duration is related to subjective sleep quality and subjective well-being: an actigraphy study. *PloS one* (8:8), e71292.

Liang, Z., Martell, M. A. C., & Nishimura, T. (2016). A personalized approach for detecting unusual sleep from time series sleep-tracking data. in *2016 IEEE International Conference on Healthcare Informatics (ICHI),* IEEE, 18–23.

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized medicine* (8:2), 161–173.

Luyster, F. S., Strollo, P. J., Zee, P. C., & Walsh, J. K. (2012). Sleep: a health imperative. *Sleep* (35:6), 727–734.

MacQueen, J. (1967). Classification and analysis of multivariate observations. in *5th Berkeley Symp. Math. Statist. Probability,* University of California Los Angeles LA USA, 281–297.

Mairs, L. & Mullan, B. (2015). Self-monitoring vs. implementation intentions: a comparison of behaviour change techniques to improve sleep hygiene and sleep outcomes in students. *International journal of behavioral medicine* (22), 635–644.

Means, M. K., Edinger, J. D., Glenn, D. M., & Fins, A. I. (2003). Accuracy of sleep perceptions among insomnia sufferers and normal sleepers. *Sleep medicine* (4:4), 285–296.

Morin, C. M., Belleville, G., Bélanger, L., & Ivers, H. (2011). The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. *Sleep* (34:5), 601–608.

Myllymäki, T., Rusko, H., Syväoja, H., Juuti, T., Kinnunen, M.-L., & Kyröläinen, H. (2012). Effects of exercise intensity and duration on nocturnal heart rate variability and sleep quality. *European journal of applied physiology* (112:3), 801–809.

National Research Council (2011). *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*.

O'Callaghan, F., Muurlink, O., & Reid, N. (2018). Effects of caffeine on sleep quality and daytime functioning. *Risk management and healthcare policy* (11), 263.

Orzeł-Gryglewska, J. (2010). Consequences of sleep deprivation. *International journal of occupational medicine and environmental health*.

Óskarsdóttir, M., Islind, A. S., August, E., Arnardóttir, E. S., Patou, F., Maier, A. M., et al. (2022). Importance of Getting Enough Sleep and Daily Activity Data to Assess Variability: Longitudinal Observational Study. *JMIR Formative Research* (6:2), e31807.

Petrakaki, D. (2017). Producing communal health through self care: the emergence of digital patient activism. *Proceedings of the 25th European Conference on Information Systems (ECIS), Guimarães, Portugal, June 5-10*, 815–827.

Pinto, L. R., Pinto, M. C. R., Goulart, L. I., Truksinas, E., Rossi, M. V., Morin, C. M., & Tufik, S. (2009). Sleep perception in insomniacs, sleep-disordered breathing patients, and healthy volunteers–an important biologic parameter of sleep. *Sleep Medicine* (10:8), 865–868.

Rezaie, L., Fobian, A. D., McCall, W. V., & Khazaie, H. (2018). Paradoxical insomnia and subjective–objective sleep discrepancy: A review. *Sleep medicine reviews* (40), 196–202.

Roenneberg, T., Wirz-Justice, A., & Merrow, M. (2003). Life between clocks: daily temporal patterns of human chronotypes. *Journal of biological rhythms* (18:1), 80–90.

Rupp, T. L. & Balkin, T. J. (2011). Comparison of Motionlogger Watch and Actiwatch actigraphs to polysomnography for sleep/wake estimation in healthy young adults. *Behavior research methods* (43), 1152–1160.

Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: an update. *Sleep Medicine Reviews* (15:4), 259–267.

Schmitz, L., Sveinbjarnarson, B. F., Gunnarsson, G. N., Davidsson, Ó. A., Davidsson, Þ. B., Arnardottir, E. S., Óskarsdóttir, M., & Islind, A. S. (2022). Towards a Digital Sleep Diary Standard. *Proceedings of the Americas Conference on Information Systems (AMCIS), Minneappolis, August 9-13*.

Sigurðardóttir, S. G., Islind, A. S., & Óskarsdóttir, M. (2022). "Collecting Data from a Mobile App and a Smartwatch Supports Treatment of Schizophrenia and Bipolar Disorder," in *Challenges of Trustable AI and Added-Value on Health,* IOS Press, pp. 239–243.

Sultan, N. (2015). Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education. *International Journal of Information Management* (35:5).

Sveinbjarnarson, B. F., Schmitz, L., Arnardottir, E. S., & Islind, A. S. (2023). The Sleep Revolution Platform: a Dynamic Data Source Pipeline and Digital Platform Architecture for Complex Sleep Data. *Current Sleep Medicine Reports*, 1–10.

Topol, E. J. (2014). Individualized medicine from prewomb to tomb. *Cell* (157:1), 241–253.

Vallo Hult, H., Islind, A. S., Rydenman, K., & Hällsjö Wekell, P. (2022). "Decreased Memory Bias via a Mobile Application: A Symptom Tracker to Monitor Children's Periodic Fever," in *Challenges of Trustable AI and Added-Value on Health,* IOS Press, pp. 915–919.

Walsh, N. P., Kashi, D. S., Edwards, J. P., Richmond, C., Oliver, S. J., Roberts, R., Izard, R. M., Jackson, S., & Greeves, J. P. (2022). Good perceived sleep quality protects against the raised risk of respiratory infection during sleep restriction in young adults. *Sleep*.

Zhang, L. & Zhao, Z.-X. (2007). Objective and subjective measures for sleep disorders. *Neuroscience bulletin* (23:4), 236.

Appendix D

Publication IV

Towards a Deeper Understanding of Sleep Stages through their Representation in the Latent Space of Variational Autoencoders

Luka Biedebach* Reykjavik University lukab@ru.is

Matias Rusanen* University of Eastern Finland matias.rusanen@uef.fi

Benedikt Hólm Þórðarson Reykjavik University benediktht@ru.is

Erna Sif Arnardóttir Reykjavik University ernasifa@ru.is

María Óskarsdóttir Reykjavik University mariaoskars@ru.is

Sami Nikkonen University of Eastern Finland sami.nikkonen@uef.fi

Henri Korkalainen University of Eastern Finland henri.korkalainen@uef.fi

Sami Myllymaa University of Eastern Finland sami.myllymaa@uef.fi

Juha Töyräs University of Eastern Finland juha.toyras@kuh.fi

Samu Kainulainen University of Eastern Finland samu.kainulainen@uef.fi

Timo Leppänen University of Eastern Finland timo.leppanen@uef.fi

Anna Sigridur Islind Reykjavik University islind@ru.is

Abstract

Artificial neural networks show great success in sleep stage classification, with an accuracy comparable to human scoring. While their ability to learn from labelled electroencephalography (EEG) signals is widely researched, the underlying learning processes remain unexplored. Variational autoencoders can capture the underlying meaning of data by encoding it into a low-dimensional space. Regularizing this space furthermore enables the generation of realistic representations of data from latent space samples. We aimed to show that this model is able to generate realistic sleep EEG. In addition, the generated sequences from different areas of the latent space are shown to have inherent meaning. The current results show the potential of variational autoencoders in understanding sleep EEG data from the perspective of unsupervised machine learning.

1. Introduction

During sleep, we wander through different stages, characterized by certain physiological features. These features and their temporal variation is traditionally recorded in polysomnography (PSG), which is a multi-signal sleep study based on multiple sensors. The results of the PSG outline the gold standard diagnostic method for many sleep disorders (Arnardottir, Islind, & Óskarsdóttir, 2021; Schmitz et al., 2022). One feature that varies significantly between different physiological sleep stages is the brain's electrical activity, recorded with electroencephalography (EEG). The EEG outlines a vital part of the PSG enabling scoring of sleep stages with the inclusion of eye movements and chin muscle tone (Berry et al., 2018). Currently in clinical practice, sleep technologists classify 30-second epochs of PSG recordings into five sleep stages; wakefulness (Wake), three non-rapid eye movement sleep (Stages N1, N2, and N3) and rapid eye movement (REM) sleep. The classification is done according to the rules set by the American Academy of Sleep Medicine (AASM) (Berry et al., 2018). However, the current five-stage and 30-second epochs process is a simplification that is needed to alleviate the workload of manual sleep staging, and both aspects lack a complete scientific justification (Himanen & Hasan, 2000). Therefore, the details of underlying feature variation of the complex sleep EEG recordings remains a subject of research. In this paper, we propose a method to explore the relationship between scored sleep stages and physiological sleep stages.

State-of-the-art machine learning models such as deep convolutional neural networks (CNNs) are capable of classifying sleep stages with similar reliability as sleep technologists (Perslev et al., 2021; Korkalainen et al., 2019; Phan & Mikkelsen, 2021; Fiorillo et al., 2019). This is a major achievement for sleep research in general and has the potential to reduce the manual workload in clinical practice. However, these models rely on supervised learning using labelled sleep recordings (Korkalainen et al., 2019). As a result, they express high classification accuracies but are limited to repeating the manual sleep staging which they are

trained with, in an automatic manner. In addition, the learning process and the used features are often untraceable and difficult to visualize.

Samek et al. pointed out, that due to the lack of transparency at the machine learning models, we can neither verify them nor learn from them (Samek, Wiegand, & Müller, 2017). Consequently, there is a rising demand toward explainable artificial intelligence (XAI) (Gerlings, Shollo, & Constantiou, 2020), i.e. machine learning models that not only provide an output but also enable the understanding how the output was achieved (Shaban-Nejad, Michalowski, Brownstein, & Buckeridge, 2021; Linardatos, Papastefanopoulos, & Kotsiantis, 2021; Rudin, 2019). There is a long withstanding discussion of the need for unpacking technology (Orlikowski, Iacono, et al., 2001; Kallinikos, 2002), and more specifically, on unpacking artificial intelligence (AI) and moving away from the black-box mentality (Castelvecchi, 2016).

In this paper, we aim to take a first step towards making machine learning models in sleep research more traceable, which is strongly needed especially in the healthcare sector. In sleep, just like in any other medical application of machine learning, the decisions made by a model come with a high responsibility, as they directly affect the health of a patient. For this reason, XAI helps medical professionals to gain trust and increase the actual usage of those systems (Xie, Gao, & Chen, 2019). Lately, generative machine learning models have been helpful in XAI through visualizations (Kahng, Thorat, Chau, Viégas, & Wattenberg, 2019). One of the most studied type of generative models is a Variational Autoencoder (VAE), a specifically structured generative autoencoder (Kingma & Welling, 2019). VAEs have been used for example to generate interpretable features of electrocardiography (ECG) (Kuznetsov, Moskalenko, Gribanov, & Zolotykh, 2021). Moreover, VAEs have increased the classification accuracy of EEG-based speech recognition systems (Krishna, Co, Carnahan, & Tewfik, 2020).

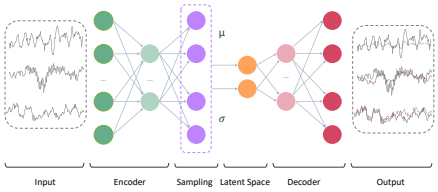


Figure 1. General Architecture of a Variational Autoencoder

Based on the previous findings, we hypothesize that VAEs have the potential to learn the underlying feature variations of sleep EEG recordings. In this proof of concept study, we aim to generate realistic sleep EEG using VAEs. In addition, we aim to show that VAEs can make an interpretable latent space using sleep EEG inputs. We furthermore discuss how this method could pave the way for a deeper understanding of sleep stages.

2. Related Work

Previously, autoencoders have been used in the context of sleep staging as a preprocessing step or as an unsupervised classifier. Najdi et al. used an autoencoder to learn a compact feature vector of PSG data in a sleep stage classification algorithm (Najdi, Gharbali, & Fonseca, 2017). Moreover, Perslev et al. utilized a typical architecture of convolutional autoencoders in another supervised sleep stage classification model (Perslev et al., 2021). Similarly, Prabhudesai et al. developed a method to automatically learn features from the raw EEG data with an autoencoder, which were then used to cluster the data to different sleep stages (Prabhudesai, Collins, & Mainsah, 2019). Autoencoders can also be used for unsupervised pre-training before supervised classification as shown by Wei et al. (Wei, Zhang, Wang, & Dang, 2018). In this study, we do not want to outperform the previous models in terms of classification accuracy but rather deepen the understanding of the sleep stages by investigating the properties learned by the autoencoder.

Variational autoencoders have shown their ability to create a meaningful latent space in other domains such as language processing (Song, Sun, Chen, Peng, & Song, 2019), image generation (Razavi, Van den Oord, & Vinyals, 2019), and cancer diagnosis (Way & Greene, 2018). In the medical field, VAEs are used to gain an understanding of ECG data (Kuznetsov et al., 2021). VAEs can also be used for emotion recognition based on EEG (Li et al., 2020) and extracting features for speech recognition on EEG data (Krishna, Tran, Carnahan, & Tewfik, 2020). To the best of our knowledge, VAEs have not been applied to sleep EEG data before as a generative model.

3. Theoretical Background

3.1. Variational Autoencoders

Autoencoders are artificial neural networks, which encode the data into a latent space and then decode it as closely as possible back into its original shape. It is a reconstruction-based form of representation learning, since the model is trained by comparing the reconstructed output with the original input (Bengio, Courville, & Vincent, 2013). The fundamental

concept of autoencoding lies within the autoencoder's architecture, consisting of an encoding function, an intermediary latent space, and a decoding function as illustrated in Figure 1 (Goodfellow, Bengio, Courville, & Bengio, 2016).

VAEs make an addition to this architecture by adding a probabilistic manipulation to the latent space variables (Kingma & Welling, 2013). In VAEs, the encoder's architecture comprises two fully connected layers connected into two latent vectors. The vectors' elements represent the mean and variance of a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ for each latent space dimension. Furthermore, the encoder comprises a sampling layer, which maps the measures of the probability distribution into the final latent space samples. These samples also compose the input of the decoder (Spinner, Körner, Görtler, & Deussen, 2018). In contrast to normal autoencoders, VAEs also function as generative models. The generative nature of VAEs emerges from the sampling layer, which enables the sampling of the probabilistic latent space, as well as from the decoder, that can be used to generate reconstructions from the latent space samples (Kingma & Welling, 2013).

Although the latent space has a simple probabilistic nature, a reconstruction loss-based optimization alone can lead to an overly complex, non-continuous, and unorganized latent space structure. In this case, generated representations of the latent space can be hard to interpret or completely unrealistic (Kingma & Welling, 2013). Therefore, VAEs introduce a regularization term in the total loss of the model. This term is added to the reconstruction loss and controls the structure of the latent space during optimization. The total loss is therefore a combination of two parts, i.e.

$$\text{Total loss} = \text{reconstruction loss} + \text{regularization},$$

where reconstruction loss is usually the mean squared error (MSE) or mean absolute error (MAE) between the input and the output of VAE for one-dimensional signals (Kuznetsov et al., 2021; Krishna, Tran, et al., 2020).

In the case of VAEs, the regularization term is defined using Kullback-Leibler (KL) divergence, which is a statistical distance measure between two distributions (Kullback & Leibler, 1951). The distance is computed in each iteration of weight optimization between distributions of the latent space samples $X \sim \mathcal{N}(\mu, \sigma^2)$ and a unit normal distribution $I \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Thus, by minimizing the KL divergence, we force the latent probability distributions to follow a normal distribution, making the latent space more organized and continuous (Kingma & Welling, 2013). The total loss can be written as follows:

$$\begin{aligned}\text{Total loss} = \text{MSE}(\text{input}, \text{reconstruction}) \\ +\text{KL}(X, I).\end{aligned}$$

Because of the difference in dimensionality between the input data and the latent space, the MSE and KL loss are averaged before summation. The additional regularization enables the creation of a continuous and potentially meaningful latent space, but reduces the autoencoder's ability to accurately create reconstructions (Asperti & Trentin, 2020). Nevertheless, we can adjust the balance between good reconstructions and more continuous latent space (Alemi et al., 2018). However, better reconstructions come with the cost of possibly overlapping latent space clusters and noisier encodings. One of the methods used in balancing between these two factors is called β-VAE (Higgins et al., 2017). This method multiplies the KL loss with a constant β. It has also been shown that monotonically or cyclically updating the β value increases the performance of VAEs as well as helps with an easily vanishing KL term (Fu et al., 2019).

As explained, the latent space of the VAE is more continuous in contrast to the sparse latent space created in normal autoencoders. As the latent space is distributed around the origin and generally shares a similar value range, a valid output can be generated from decoding points in the latent space (Spinner et al., 2018). Due to these special properties of the latent space in variational autoencoders, the newly generated samples and their position in the latent space become interpretable.

3.2. Convolutional Layers

The advantage of convolutional neural networks (CNN) is that they extract visually meaningful information. Even though CNNs are most commonly used for image processing, they have also shown to be a suitable approach for transforming EEG data (Bashivan, Rish, Yeasin, & Codella, 2015). A convolutional layer of a CNN slides a kernel of a filter over the input to extract features at each position. A filter is therefore a stack of matrices, the kernels, which factors are learned during training (O'Shea & Nash, 2015). The kernel size defines the size of the sliding window which is passed over the data. Smaller kernels tend to collect more local information, while larger kernels extract the global, high-level features (Gu et al., 2018). CNNs usually comprise multiple convolutional layers with a different number of filters and different kernel sizes. In this way, the architecture of the CNN is constructed to extract information on multiple scales. Furthermore,

using convolutional layers in VAEs, the size of the input can be gradually decreased towards the latent space to reduce dimensionality while extracting information.

4. Method

4.1. Data

For this paper, we used 50 PSG recordings, which totals in 381.13 hours of EEG data. The data collection was approved by the National Bioethics Committee of Iceland (21-070). Informed written consent was obtained from all participants before measurements. We have a diverse study population with 27 male, 19 female, and 4 unspecified-gendered participants. The study population included participants with and without diagnosed sleep disorders. More information about the study population can be found in Table 1.

Table 1. Demographic information of the study population (n=50)

Variable	Mean ± SD
Age [years]	44.2 ± 13.4
Weight [kg]	84.1 ± 21.7
Height [cm]	174.9 ± 9.9
BMI [kg/m^2]	27.3 ± 5.3
AHI [1/h]	12.0 ± 13.2

SD = standard deviation, BMI = body mass index, AHI = apnea-hypopnea index

The PSG recordings were conducted at Reykjavik University as part of the Sleep Revolution project. The PSG was set up by a professional sleep technologists and the participants slept at home in their natural sleeping environment. The Type II PSG recordings were conducted using a portable PSG device (Nox A1, Nox Medical, Reykjavik, Iceland) and included EEG as recommended by the AASM (Berry et al., 2018). We used the F4-M1 channel from the EEG recordings as a single-channel input to the VAE, as it is commonly used in manual sleep staging. Only one channel was used to keep the feature variation of the input EEGs reasonable. For visualization and exploration of the latent space, we used the manual scoring of sleep stages, which was conducted by an experienced sleep professional according to the AASM scoring manual (Berry et al., 2018).

4.2. Preprocessing

The EEG signals were originally saved using 200 Hz sampling frequency in the Noxturnal (Nox Medical) software and exported to EDF format. The signals were then preprocessed with Python according to the following steps. First, we downsampled the signals to 64 Hz to reduce the computational burden and complexity of EEG signals. Second, we applied high-pass filters with a cut-off frequency of 0.3 Hz, as recommended in the AASM scoring manual (Berry et al., 2018). Finally, we scaled the signal amplitudes into a range between 0 and 1 using min-max scaling. We confirmed that the EEG signals appeared normal after each preprocessing step as illustrated in Figure 2. These preprocessing steps were conducted per subject to preserve the amplitude variation in each recording.

The recording has a length of approximately 7 hours per participant. To work with this data in a machine learning context, we split it into smaller 10-second sub-sequences. Sleep stages were manually scored in 30 second windows, but we chose the time window of 10 seconds to reduce the length of the time series processed by the VAE. The 10-second segments were randomly divided into training (90%) and testing (10%) sets, resulting in 357.9 hours or 128857 segments of EEG data for training and 39.8 hours or 14318 segments of EEG data for testing. In unsupervised machine learning the division to train and test sets is not mandatory but we chose to include it for experimental reasons.

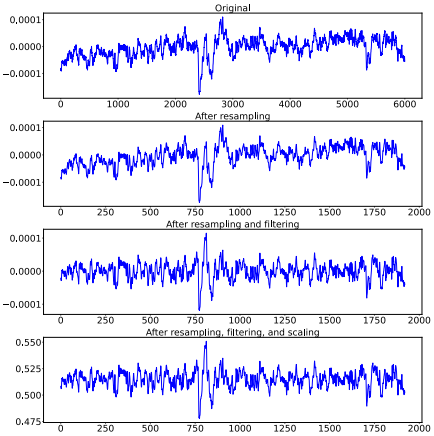


Figure 2. A 30-second sample of the input EEG after each of the preprocessing steps.

4.3. Optimization

The models were implemented using TensorFlow version 2.8.0 (Abadi et al., 2015) and Keras application programming interface. We optimized the VAEs using the Adaptive Moment Estimation (Adam) algorithm (Kingma & Ba, 2015) with default Keras configurations.

The weights were updated in batches of size 100 until the total loss converged or until 100 training epochs.

During the training of the VAE, both the reconstruction error as well as the KL divergence were taken into account. A common problem with VAEs is the KL divergence collapse problem (Asperti & Trentin, 2020; Alemi et al., 2018), which arises from unequal scales of the reconstruction error and the KL divergence. To ensure a balance between them, we used the β-VAE method.

5. Experiments

5.1. Dense VAE

As a proof-of-concept of VAEs operating with sleep EEG data, we ran experiments on the most simple VAE architecture. This architecture comprised only a single dense connection between the input and latent space parameters as well as the latent space sample and the output of the decoder. Furthermore, we used the input size of the signal (640 samples) as the latent space dimension to further increase the simplicity of our method. We increased the weight of reconstruction loss in the total loss using constant $\beta = 0.0001$ multiplying the KL term. The learning rate was decreased from 0.01 with 0.001 steps after each iteration until optimization stopped or the learning rate reached a value of 0.001.

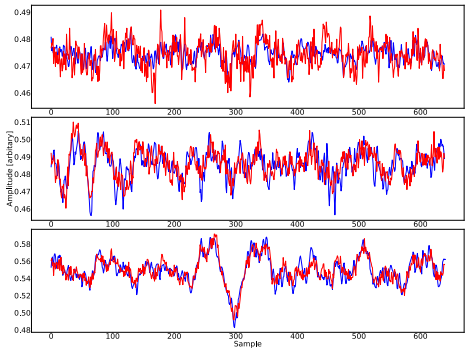


Figure 3. Three exemplary inputs of 10-second EEG segments (blue) and their reconstructions (red) using the dense model.

Our experiment to reconstruct EEG with a simple dense VAE clearly showed the ability of VAEs to work with highly complex EEG inputs. The reconstructions shown in Figure 3 were achieved after 100 epochs of training. Despite the desired reconstructions, this model was unsuitable for the intention to explore the sleep EEG data through latent space, as the latent space had no dimension reduction relative to the input data.

5.2. CNN VAE

Moving from high-dimensional latent space to reducing the dimensions into something that can be visualized, we chose to experiment with three dimensions in the latent space. Following the major change in the dimensional reduction of input data, the purpose was not to reach similar reconstructions as shown with our simple dense model. Instead, we experimented with whether the VAE can still extract features relative to input data and generate realistic EEG samples. For meaningful feature extraction, we included a CNN layer in the VAE's architecture. Keeping the experiment simple, only one convolutional or respectively deconvolutional layer was added to both the encoder and decoder. We used 256 filters with kernels of size 5 for the convolutions. Then, we reduced the length of the sequence with max-pooling. In the following layer, we flattened the sequence into one dimension, before connecting it to the latent vectors. Here, instead of forwarding the exact position in the latent space to the decoder, the mean and variance of a normal distribution were used to sample a position in the latent space. At this point the data was compressed to a vector of length three, which was then transformed back into its original shape by the following layers.

In the decoder, we used a dense layer which mirrored the transformation of the sampling and a reshape layer that mirrored the transformation of the flattening layer in the encoder. Then, an up-sampling layer was used to mirror the max-pooling. Finally, a deconvolutional layer with one filter and kernel size 5 brought the data back into their original shape. Both in the encoder and in the decoder, we used Rectified Linear Units (ReLU) as activation functions. For the optimization of this model, we gradually increased the weight of the KL divergence from $\beta = 0.01$ to $\beta = 1$ in 100 epochs. In this manner, the model should first learn the reconstructions, after which the latent space is made regular (Higgins et al., 2017). A constant learning rate of 0.001 was used for optimization of this model. In the following Results section, we refer to the results achieved with the CNN VAE.

6. Evaluation

6.1. Turing Test

The Turing test is an experimental set-up developed by Alan Turing to test the intelligence of a machine (Turing, 2009). Originally, the test was designed to

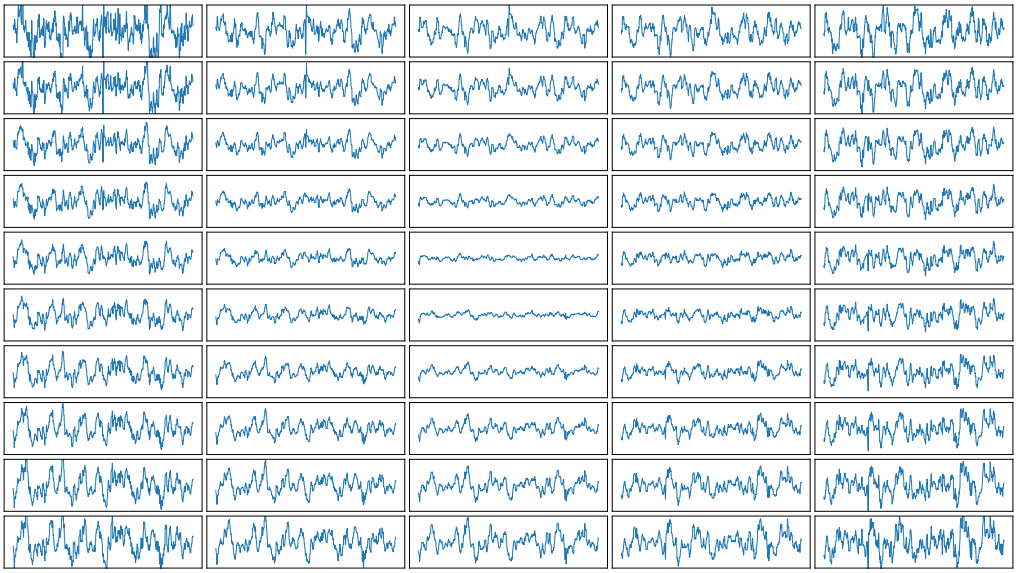


Figure 4. Generated artificial EEG segments (10 seconds) sampled with linear intervals from the first and the third axis of the latent space, keeping the second dimension coordinate constantly as zero.

determine whether an interrogator can distinguish a human from an AI in a dialogue with both of them. We used the principles of this test by confronting a sleep technologist with both real and artificially created EEG sequences. This way, we evaluated whether the sequences generated by our VAE were realistic.

We sampled 10-second segments randomly from the input EEG signals. In addition, we created artificial EEG segments by randomly sampling each of the latent space coordinates from a uniform distribution between -3 and 3 and passing the resulting point to the decoder. In the first step, the totalling 50 signal segments were then distributed randomly on a 5x10 grid including 46 real and four artificial EEG sequences. A sleep expert was asked to point the four artificial EEG sequences out. In a second step, we confronted the sleep technologist with a 3x6 grid (18 sequences) including likewise four artificially created EEG sequences.

6.2. Manual Review

In order to verify that our model could not only generate realistic EEG sequences, but also created a meaningful latent space, we manually reviewed the generated sequences with the sleep technologist. In this experiment, we showed the sleep technologist two maps of artificial EEG segments on a 5x10 grid sampled and decoded from the latent space. One map showed samples from the first and second axis of the latent space, while a second map showed the first and third axis. In both maps, the excluded dimension was kept at a constant value of zero. For visualisation purposes, we extracted the mean from each segment and fixed the y-axis limits of the subplots to be constant. To gain an understanding of different axes, sleep technologist was asked to give an estimation of the sleep stage of different EEG sequence. We furthermore asked for the presence of sleep stage specific patterns and artifacts.

7. Results

The Turing test-like experiment showed that the sleep technologist was not able to identify any of the four artificial EEG sequences from the real examples in a set-up of 50 sequences. Also, in a set-up of 18 sequences, the sleep technologist was unable to identify the four artificial ones. From this, we conclude that our VAE can generate realistically looking EEG sequences.

Using the CNN VAE, the generated artificial EEG sequences showed different features according to the latent space position they originated from. Figure 4 shows a map of points sampled from the first and third axis of the latent space, keeping the second axis coordinate constant at zero. The variation of the

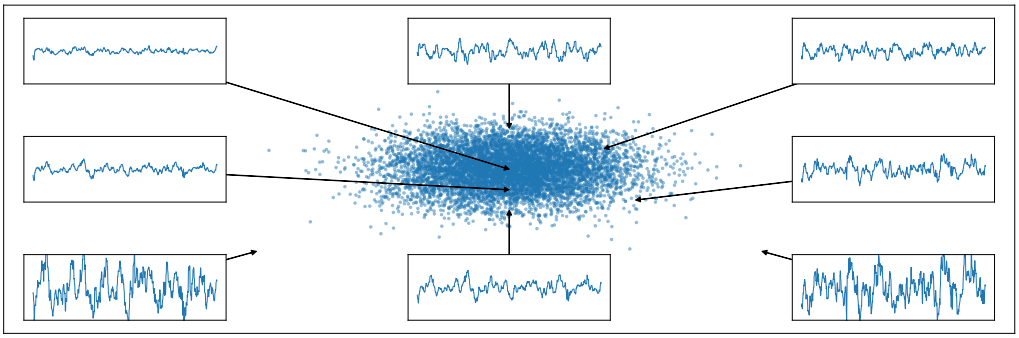


Figure 5. Signal segments (10 seconds) sampled from the latent space axes 0, and 2.

sequences sampled from different positions were clearly visible. We can observe lower amplitude signals to originate from the center of the latent space while amplitudes increased when increasing values of the first and third axis. In some parts of the latent space the VAE also generated sequences that do not look like EEG at all or resemble artifacts. These sequences might arise due to noncontinuous area of latent space or might be learned from artifacts in the training data.

The sleep technologist confirmed by manual review that samples from certain positions along axes of the latent space resemble certain sleep stages. It needs to be noted, that this was not a proper scoring according to the AASM rules, but instead a subjective estimation based solely on the shape of the signal. However, this variation showed that the samples were not randomly generated, and that the axis of the latent space contains meaning and reflects features typical for different sleep stages. Figure 5 shows samples generated from exemplary positions in the latent space. In this visualization, the second axis was held at a constant value of 0. The sequence in the top left corner was perceived as REM or N1 sleep by the sleep technologist, while the sample in the bottom left corner was perceived as deep sleep (N3).

Figure 6 shows clustering of the training data along the first axis of the latent space. In the three-dimensional visualization of the latent space, the sleep stages were slightly organized into clusters. However, depending on which axis was visualized, more clusters not related to the sleep stages become visible.

8. Discussion

In this paper, we aimed to show that VAEs can be applied to sleep EEG data. This is an important contribution to the field of information technology in sleep research as this model proposes novel methods to generate insights into sleep structure. The main novelty of this work is that it follows the principles of XAI, by making the latent space interpretable and the learning process traceable. Secondly, the proposed model is fully unsupervised, and hence does not require any manually scored data nor carry the bias introduced by manual scoring during the training.

The present results indicate that VAEs are able to generate realistic but synthetic samples of EEG with varying features that are common to sleep EEG data. In addition, we showed that the created latent space was not random, but reflected features of different sleep stages in different positions along the axes. Finally, the results illustrate that a simple convolutional VAE was capable of generating preliminary clustering of the EEG data in the latent space. Therefore, we suggest that this method could open a way to understand sleep stages in a more sophisticated manner than previously achieved through manual analysis and unpack some of the criticized mystery related to AI. It also shows preliminary potential of comparing unsupervised sleep staging to supervised sleep staging and manual sleep staging through labeling of the latent clusters.

The artificial EEG sequences generated with the VAE were not distinguishable from real EEG segments to an expert sleep technologist. Although this observation contains the bias of a subjective opinion of a single sleep technologist, it shows that the artificial sequences can be considered realistic. It furthermore indicates that even the simple neural network model can learn some features representing the input data. It should also be noted, that the method is scalable to studying neural networks of different architectures in the encoder and decoder. This method could therefore help in understanding some of the already existing sleep staging models (Perslev et al., 2021; Korkalainen et al.,

2019; Phan & Mikkelsen, 2021; Fiorillo et al., 2019).

The variation within our latent space, representing features that might be attributed to different sleep stages, shows how the model built an understanding of the EEG sequences in an interpretable way. The high complexity inherent in the combination of thousands of weights within the neural network prevents us from fully tracing what is learned during the training. However, the continuous latent space created by the VAE is a sophisticated visual approach to understand the features that are learned by the model. In order to get a more holistic evaluation of the methodology, it would require an in-depth analysis by multiple sleep technologists. That is however outside the scope of this paper. In addition, more quantitative analyses of the generated EEG signals are needed in the future.

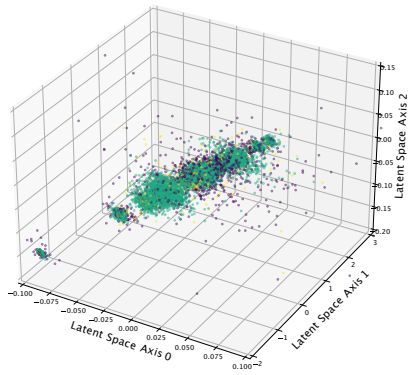


Figure 6. Embedding of EEG sequences used in training within the three-dimensional latent space, colored by sleep stage. Values on the axes represent means of the latent space samples. Colors: Yellow = REM, Green = N3, and Purple = Wake.

Compressing the data into a three-dimensional and continuous latent space comes with a certain cost. First, the low dimensionality creates a tight bottleneck in which inevitably information is lost. Second, the sampling layer introduces randomness to the model. Hence, the autoencoder faces a trade-off between accurate reconstructions of the input signals (as seen in the dense model) and a meaningful latent space suitable for generating data (as seen in the CNN model). Regarding the high complexity of our input signal, a sequence of a multi-frequency biosignal with a length of 640 samples, it is hard to achieve accurate reconstructions even with the dense model that operated without any dimensionality reduction. However, the purpose of the model is not to perfectly reconstruct compressed input signals, but to create a latent space from which realistic EEG can be generated. From this perspective, the reconstructions are not an issue, as the individual peaks and troughs are irrelevant for realistic EEGs, while the general frequency and recurring patterns matter more (Berry et al., 2018).

Moving away from manual review, we can also perceive an irregular distribution of sleep stages in the latent space when using the manual sleep stage scoring as labels. The slight clustering of sleep stages within the latent space hints towards further possibilities for unsupervised sleep staging. However, to achieve this, more sophisticated models that better capture the feature variation of input EEGs are likely needed. One possibility could be a concatenated model with separate branches of convolutional and dense layers, as proposed in (Kuznetsov et al., 2021). However, adding complexity to the model might make the optimization of the model more difficult. This in turn highlights the need for more adaptive optimization methods such as VAEs with calibrated decoder (Rybkin, Daniilidis, & Levine, 2020) or VAEs utilizing single-parameter, continuous Bernoulli distributions (Loaiza-Ganem & Cunningham, 2019). Furthermore, some clusters we observed in the latent space, which were not related to sleep stages might represent other factors, e.g. patient demographics. In order to study this assumption, other attributes such as age, gender, and sleeping disorders need be considered in the future. Another method could be splitting the data before training into subgroups based on other attributes and comparing the latent spaces that are created. Especially training one model on recordings by participants with obstructive sleep apnea (OSA) and another model on participants without any sleep disorders could reveal differences in their EEG features. Before these experiments are possible, more methodological studies on using VAE with sleep EEG data need to be conducted.

9. Conclusion

We can conclude, that this paper is preliminary work that explored the possibilities of VAEs in sleep research, opening up several new research directions in the future. This study contributes an addition to traditional machine learning-assisted sleep research in the following ways: i) by introducing a method for generating realistic artificial EEG; ii) by showing potential of providing in-depth understanding of sleep EEG and sleep staging through XAI, and; iii) by creating the foundation for attempting unsupervised sleep staging through clustering in the latent space. Our findings are relevant for the field of sleep research and health information systems in general because we show how a VAE can act as a generative and

interpretable model for EEG data. Generating realistic EEG sequences is not only relevant for sleep research but can also be used as a method in various medical domains, and as such, apply to a variety of health information systems issues. We hope that introducing XAI in sleep research could increase the acceptance and usage of AI systems by sleep professionals in the hospital and beyond.

10. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965417. We thank Sigríður Sigurðardóttir for the manual review of the generated EEG sequences. The first two authors Luka Biedebach and Matias Rusanen contributed equally and share the first authorship. The senior author of this paper is Anna Sigridur Islind.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* (Software available from tensorflow.org)

Alemi, A. A., Poole, B., Fische, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2018). Fixing a broken ELBO. *35th International Conference on Machine Learning, ICML 2018*, *1*, 245–265.

Arnardottir, E. S., Islind, A. S., & Óskarsdóttir, M. (2021). The future of sleep measurements: A review and perspective. *Sleep medicine clinics*, *16*(3), 447–464.

Asperti, A., & Trentin, M. (2020). Balancing reconstruction error and Kullback-Leibler divergence in variational autoencoders. *IEEE Access*, *8*(1), 199440–199448.

Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2015). Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Lloyd, R. M., Quan, S. F., ... Vaughn, B. V. (2018). *AASM Manual for the Scoring of Sleep and Associated Events* (Tech. Rep.). Amer. Acad. Sleep Med., Darien, IL, USA: American Academy of Sleep Medicine.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, *538*(7623), 20.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P. L., Favaro, P., Roth, C., ... Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, *48*, 101204.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., & Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the need for explainable artificial intelligence (xai). *arXiv preprint arXiv:2012.01007*.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT Press Cambridge.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... others (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework. In *Iclr* (Vol. 44).

Himanen, S.-L., & Hasan, J. (2000). Limitations of rechtschaffen and kales. *Sleep medicine reviews*, *4*(2), 149–167.

Kahng, M., Thorat, N., Chau, D. H. P., Viégas, F. B., & Wattenberg, M. (2019). GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 310–320.

Kallinikos, J. (2002). Reopening the black box of technology artifacts and human agency.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, *12*(4), 307–392.

Korkalainen, H., Leppanen, T., Aakko, J., Nikkonen, S., Kainulainen, S., Leino, A., ... Toyras, J. (2019). Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 2073–2081.

Krishna, G., Co, T., Carnahan, M., & Tewfik, A. H. (2020). Constrained Variational Autoencoder

for improving EEG based Speech Recognition Systems.

Krishna, G., Tran, C., Carnahan, M., & Tewfik, A. (2020). *Constrained variational autoencoder for improving eeg based speech recognition systems.* arXiv.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Kuznetsov, V., Moskalenko, V., Gribanov, D., & Zolotykh, N. Y. (2021). Interpretable feature generation in ecg using a variational autoencoder. *Frontiers in genetics*, *12*.

Li, X., Zhao, Z., Song, D., Zhang, Y., Pan, J., Wu, L., … Wang, D. (2020). Latent factor decoding of multi-channel eeg for emotion recognition through autoencoder-like neural networks. *Frontiers in Neuroscience*, *14*.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1).

Loaiza-Ganem, G., & Cunningham, J. P. (2019). The continuous bernoulli: Fixing a pervasive error in variational autoencoders. *Advances in Neural Information Processing Systems*, *32*(NeurIPS), 1–11.

Najdi, S., Gharbali, A. A., & Fonseca, J. M. (2017). Feature transformation based on stacked sparse autoencoders for sleep stage classification. In *Doctoral conference on computing, electrical and industrial systems* (pp. 191–200).

Orlikowski, W. J., Iacono, C. S., et al. (2001). Desperately seeking the "it" in it research–a call to theorizing the it artifact. *Information systems research*, *12*(2), 121–134.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*, *4*(1), 1–12.

Phan, H., & Mikkelsen, K. (2021). *Automatic Sleep Staging: Recent Development, Challenges, and Future Directions.*

Prabhudesai, K. S., Collins, L. M., & Mainsah, B. O. (2019). Automated feature learning using deep convolutional auto-encoder neural network for clustering electroencephalograms into sleep stages. In *2019 9th international ieee/embs conference on neural engineering (ner)* (pp. 937–940).

Razavi, A., Van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, *32*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Rybkin, O., Daniilidis, K., & Levine, S. (2020). Simple and effective VAE training with calibrated decoders.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Schmitz, L., Sveinbjarnarson, B. F., Gunnarsson, G. N., Davidsson, O. A., Davidsson, T. B., Arnardóttir, E. S., … Islind, A. S. (2022). Towards a digital sleep diary standard. In *Americas conference on information systems.*

Shaban-Nejad, A., Michalowski, M., Brownstein, J., & Buckeridge, D. (2021). Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare. *IEEE Journal of Biomedical and Health Informatics*, *25*(7), 2374–2375.

Song, T., Sun, J., Chen, B., Peng, W., & Song, J. (2019). Latent space expanded variational autoencoder for sentence generation. *IEEE Access*, *7*, 144618–144627.

Spinner, T., Körner, J., Görtler, J., & Deussen, O. (2018). Towards an interpretable latent space : an intuitive comparison of autoencoders with variational autoencoders. In *Proceedings of the workshop on visualization for ai explainability 2018 (visxai).*

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the turing test* (pp. 23–65). Springer.

Way, G. P., & Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *Pacific symposium on biocomputing 2018: Proceedings of the pacific symposium* (pp. 80–91).

Wei, R., Zhang, X., Wang, J., & Dang, X. (2018). The research of sleep staging based on single-lead electrocardiogram and deep neural network. *Biomedical engineering letters*, *8*(1), 87–93.

Xie, Y., Gao, G., & Chen, X. (2019). Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv:1902.06019*.

Appendix E

Publication V

Deriving Association Rules from User Engagement in a Digital Therapeutics Application for Sleep Improvement

Abstract

Background: The demand for sleep interventions is high and steadily growing. Digital therapeutics (DTx) can be a way to tackle this challenge for at least a portion of the patients. Mobile applications can help individuals improve their sleep remotely, over an extended time, and with less effort from medical professionals. The severity of obstructive sleep apnea (OSA), as one of the most prevalent and consequential sleep disorders, can be reduced with health-supporting behavioral changes such as physical exercise and weight loss and, therefore, acts as a promising application for DTx.

Objective: We aimed to analyze a digital intervention from a medical and technological perspective by moving beyond clinical markers and exploring deeper how the DTx application was used and how that may be related to the study outcome. We introduced a way of using unsupervised machine learning to analyze the participants' sleep, behavior, and engagement with the DTx application on a day-to-day level.

Methods: A lifestyle intervention study (n=55) targeted at adults with mild-to-moderate OSA aimed to reduce their OSA severity using a DTx application over a study period of 12 weeks. The participants' OSA severity was assessed through a polysomnography at the beginning and at the end of the study period and the participants tracked their sleep with a digital sleep diary and a smartwatch over the course of the entire study. The DTx application furthermore provided data on when and how the participants pursued the proposed lifestyle interventions. This multimodal data was explored through descriptive statistics, and association rules were derived using the apriori algorithm.

Results: Analyzing the interaction of the participants with the application showed which lifestyle interventions they pursued and how their behavior and sleep changed over time. The participants with reduced OSA severity showed higher engagement with the DTx application, particularly the food-related interventions. The association rules showed that changes in awakenings during the night, staying awake in bed in the morning, and sleep quality frequently co-occurred with the education and movement missions.

Conclusion: The study showed that DTx can be an effective treatment approach for some participants, particularly those who showed active engagement with the DTx application. We furthermore showed the richness of the different data sources offered in a digital intervention using wearables and how they can be employed to get an in-depth understanding of the study.

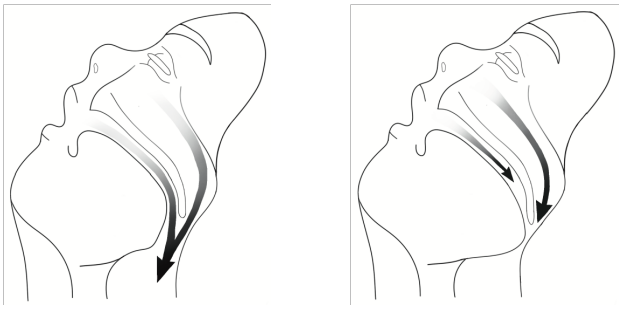
Keywords: Sleep, DTx, Wearables, Association Rules, Lifestyle Intervention.

Introduction

Digital Therapeutics (DTx) are expected to influence the way healthcare is delivered all around the world [10]. Wang et al. define DTx as *software that provides evidence-based medical interventions for disease or disorder prevention, management, and treatment* [36] and DTx is a central element within digital health. This three-fold focus on preventing, managing, and

treating diseases through DTx has shown to be successful for chronic and difficult-to-treat diseases and to lead to sustainable long-term outcomes [17]. Since sleep is one of the pillars of health [8] and obstructive sleep apnea (OSA) is highly prevalent in the general population [27], it is vital to study the relationship between DTx and OSA treatment. DTx has the strength of delivering care remotely, which enables users to take part in interventions over a continuous and extended period in the comfort of their own homes. It is a "one-to-many model of care," providing personalized healthcare while utilizing fewer resources [24]. Fürstenau et al. identify the combination with wearables, which can track sleep and activity longitudinally, as a promising future research area [15]. In this work, we will both analyze usage patterns in the DTx application enriched with information from a digital sleep diary and clinical markers from a lifestyle study for adults with OSA.

Sleep-disordered breathing encompasses a range of breathing-related sleep disorders, from habitual snoring to severe OSA. OSA is the most common form of sleep disorders worldwide, and is characterized by temporary interruptions of breathing during sleep due to upper airway obstruction [30, 33]. Globally, an estimated 425 million individuals suffer from moderate to severe OSA (defined as 15 or more events per hour) [6]. Clinically, apnea is defined as a complete cessation of airflow lasting more than 10 seconds, while hypopnea refers to a partial reduction of airflow (by at least 30%) for at least 10 seconds [27]; the Apnea-hypopnea Index (AHI) quantifies these events per hour. These disruptions frequently lead to reduced blood oxygen levels, triggering awakenings from sleep, which in turn prevents OSA patients from achieving sufficient restorative sleep. As they often return to normal breathing without recalling the interruptions, approximately 80% of individuals with OSA remain undiagnosed [29]. Figure 1 illustrates the mechanism of airway obstruction in OSA patients.



a.) A normal breath b.) An upper airway obstruction

Figure 1: A normal breath vs. an upper airway obstruction

Managing OSA effectively remains challenging despite the availability of efficacious therapies. The most common treatment for OSA is continuous positive airway pressure (CPAP) therapy, which applies a constant level of pressure to the upper respiratory tract through a nasal or oronasal mask. Even though CPAP can dramatically reduce OSA events and improve symptoms, ensuring long-term adherence to CPAP is challenging in clinical practice. Many OSA patients struggle with CPAP due to discomfort, inconvenience, or lack of perceived benefit, undermining its real-world effectiveness. Contemporary research aims to improve OSA without medical devices but rather by initiating health-supporting behavioral changes, e.g., an exercise-based lifestyle intervention [12], reducing obesity levels [13], or improving muscle

tone with myofunctional therapy [9]. In this context, digitally delivered interventions for OSA are a promising research direction as a means to educate patients, promote healthy behavior change, and remotely monitor progress over an extended period.

DTx represents a convergence of medical science, interaction design, and computer science, with the aim to transform healthcare in general and chronic disease management in particular. Therefore, when testing their effectiveness on participants, the application should be viewed from different perspectives, including known clinical markers of treatment success as well as insights from user engagement. In this paper, we focused our efforts on answering the following research questions (RQs):

RQ1: *What are engagement patterns of participants in the DTx application?*

RQ2: *How is engagement related to changes in the participant's behavior and sleep?*

In the following sections, we review existing research on DTx for sleep improvement and digital therapeutics using wearables, introduce the lifestyle intervention study, describe the statistical methods used for analysis, and discuss the results in the context of existing literature on digital health and personalized medicine.

Related Work

Behavioral interventions are a common approach to improving sleep quality for people with and without sleep disorders [26]. A web-based DTx application to improve sleep efficiency showed significant improvements for people with insomnia [12]. Web-based DTx for sleep improvement has also shown positive effects on the sleep of people without insomnia [34]. Comparing different forms of DTx, Luik et al. [23] identified three levels i.) DTx as support for offline therapy, ii.) DTx as the main therapy with the support of a health professional, iii.) DTx as a fully automated digital intervention program. One example is the work by Wickwire et al. [37], who used a combination of web- and app-based DTx in parallel with a commercial off-the-shelf sleep tracker. There are no publications focusing on DTx applications for OSA treatment in the existing literature, making this research area highly relevant. Friðgeirsdóttir et al. [13] was the first randomized control study to use exercise and a DTx application to reduce OSA severity. We aimed to build upon and complement their work with in-depth analysis of the DTx intervention group.

Research has extensively demonstrated the effectiveness of digital interventions for sleep improvement, but most research has been done in the medical field. However, given the technological nature of these interventions, a comprehensive evaluation requires considering both medical and digital perspectives. Specifically, the interaction of the participant with the application is crucial for ensuring that such interventions are engaging, intuitive, and tailored to their needs. Aji et al. [3] conducted a systematic review of existing DTx applications for sleep improvement. They identified personalized sleep feedback, educational material, and a digital sleep diary as their most common features.

Understanding how participants engage with digital health interventions and whether that engagement translates into better outcomes is an active area of research in medicine and health informatics. Previous studies have approached this issue by analyzing engagement data to identify usage metrics or patterns and then examining their association with clinical or behavioral outcomes. Overall, their findings suggest that patients who actively use and participate in digital interventions are more likely to benefit from them [25, 5].

Association rule mining is used to analyze co-occurrences of events and has been applied in different fields in digital health. For example, Concaro et al. [10] used clinical variables from medical health records to analyze the relationship between clinical variables and drug effects. There are multiple publications that used association rule mining to identify patterns in the data from sleep questionnaires, sleep parameters from polysomnography (PSG), and demographic information about participants [1, 21, 17]. The publication by Liang et al. [22] shows how longitudinal data from a smartwatch can be used to discover factors relevant to people's sleep. We pursue a similar research direction by applying association rule mining on the data collected through a DTx application, identifying patterns in the engagement of the participant in the sleep lifestyle intervention.

Methods

Study Design

This research is based on a study conducted within the EU Horizon Project *Sleep Revolution* [4]. The study spanned a 12-week period and participants were asked to (i) have a three-night self-applied PSG at the beginning and at the end of the study, (ii) wear a Withings smartwatch (Issy-les-Moulineaux, France), (iii) report their sleep in a digital sleep diary mobile application and, (iv) follow a lifestyle intervention program with the aim of improving their lifestyle, and ultimately their OSA for 12 weeks.. The study is covered by ethical approval of the National Bioethics Committee of Iceland (VSN-22-082). The study population included 192 adults between the age of 18–50 years, who have a body mass index (BMI) ≥ 25 <42 kg/m^2. The gender balance was 50.9% males. The participants had mild-to-moderate OSA, with an AHI between 5 and 30. The study population was randomly divided into three intervention groups. One of these groups used a digital health program developed by Sidekick Health (Reykjavik, Iceland). The participants used the Sidekick Digital Therapeutics (SK-DTx) application as the intervention program. Friðgeirsdóttir et al. conducted a statistical analysis on all three groups of the randomized control trial [13]. Their work showed that the DTx group reduced weight, neck circumference, body fat, visceral adiposity, and skeletal muscle mass. However, there was no significant improvement in OSA severity in the DTx group. In this paper, we aim to extend their work by analyzing the participants' engagement with the DTx application in relation to their improvement in OSA severity, as well as changes in their sleep and behavior on a day-to-day level. The DTx intervention group included 55 participants with a mean age of 37.1 years and a mean BMI of 32.8 kg/m^2. They had an average AHI of 12.1, i.e., on average, 12 apneas or hypopneas per hour of sleep. An overview of the demographic information of the DTx group can be found in Table 1.

Table 1: Demographic Information

Variable	**Mean ± Standard Deviation**
Age [years]	37.1 $\pm$ 6.8
Body Mass Index [kg/m^2]	32.9 $\pm$ 4.1
Apnea-hypopnea Index [events/h]	12.1 $\pm$ 9.7

The DTx application provides the participant with suggestions for behavioral interventions to improve their sleep. The DTx application guides the participants to improve their lifestyle

by goalsetting, self-monitoring, and completion of health-related tasks. This touches on different areas of healthy lifestyle i.) creating awareness for eating habits, nutrition, and diet through educational material and meal logging, ii.) attempting to prevent stress through relaxation, meditation, and mindfulness exercises iii.) it motivates participants to do activity and exercise iv.) providing general health education to the participant, and v.) helping the participant to take care of their own health through tracking and reminders. The overall goal of the intervention is to increase the frequency of healthy lifestyle behaviors and, by that, improve the health of the participant.

The intervention concept of the application is based on *missions* that the participant is asked to follow. Missions addressing these different areas of healthy lifestyle appear on the dashboard and the participants can select which ones they want to complete. An overview of the different mission categories can be seen in Table 2. The types of missions and the content of the missions have been designed particularly for sleep improvement. However, the application can be customized to other conditions and interventions as well and has shown success, for example, in obesity [35], and diabetes [16]. The application uses gamification in terms of design, competition, and rewards. This way, the DTx application uses storytelling to frame the mission as enjoyable tasks. The tasks provide instant gratification to the participant by showing personal achievements and providing social interactions. Participants track their personal achievements by collecting points, which earn them virtual rewards. Social interaction is created through messages from a coach, with whom the participants can communicate through the app. Additionally, a mascot aims to motivate the participant to follow the interventions. These gamification elements aim to nudge the participant towards a healthier behavior [38].

Table 2: Types of Missions

Mission	Description
Education	Video, audio, and written content are shown to the participant. The educational material is tailored to the specific intervention program, and this study includes content about good habits to improve sleep, as well as general lifestyle and health education.
Clinic	Involves logging weight, blood pressure, and other physical measurements, as well as reminders for taking supplements.
Mind	The participant is asked to log their stress and energy level, as well as their quality of sleep. This category furthermore involves different breathing and meditation exercises.
Move	Participants record their physical activity, which may include anything from daily steps to structured exercise or sports.
Food	Food intake is tracked and divided into vegetables, nuts, fruit, snacks, and candy. Beverages are tracked and divided into water and soda. This category also includes alcohol and nicotine.

Data

The study set-up, including PSG, a smartwatch, a digital sleep diary and the DTx app results in a diverse data set. In the following section, we are going to explain our four main data sources, provide an overview of the size of each data set, and describe the features that are included.

Polysomnography Data

The PSG data was manually reviewed by a sleep expert according to the American Academy of Sleep Medicine (AASM) scoring manual [7]. The self-applied PSG set-up included the classical sensors, with simplified electroencephalography using only frontal electrodes [29]. The participants recorded one to three nights of sleep with this set-up before and after the intervention period. The AHI values derived from the manual scoring of these recordings (up to three nights) were averaged to one value before and one value after for each participant.

DTx Data

The data from the DTx application is organized by the missions that participants completed. Each mission record has information about the specific mission name, mission category and the completion day and time. In total, there were 92 different missions in five mission categories. Participants could complete multiple missions per day. Apart from the mission data set, we also had information about how often the participant interacted with a health coach through the app. Here, only the day of interaction and whether the participant sent or received a message is recorded. The content of the messages was not recorded.

Digital Sleep Diary Data

As part of the study, the participants used a digital sleep diary app. It was developed within the Sleep Revolution project and is based on the research of Schmitz et al. [32]. In the app, the participants were asked to report the quality of their sleep in the morning and report information about their day in the evening. The subjective sleep quality and stress were reported on a Likert scale, with one being the lowest and five the highest rating. Participants also reported how long they stayed in bed awake in the morning before getting out of bed. The participants had an average sleep quality rating of 2.14. In the evening diary, the participants were asked about their stress level, pain, screen time, exercise, and medication intake.

If the participants did not fill out the sleep diary, they were reminded using notifications. The participants filled out the morning diary for 48 days on average and the evening diary for 42 days. We merged the diary entries from the evening diary, with the entries from the morning diary on the following day. Therefore, each record in the data set represents one night of sleep.

Smartwatch Data

The smartwatch provided metrics such as sleep duration, the time participants spent awake during the night, and the number of nocturnal awakenings. These awakenings could range from seconds to several minutes or hours. The smartwatch also tracked sleep onset latency, defined as the time taken to fall asleep. Beyond sleep, the smartwatch also recorded the number of steps the participants took during the day. The participants were asked to wear the smartwatch for the entire study duration, but there were varying levels of adherence. On average, each participant wore the smart watch for 62 days over the entire study period of 12 weeks. The average sleep duration across participants was 7 hours and 33 min, with a standard deviation of 55 minutes.

Data Analysis

As a first step, the health data from these different sources was merged, joining the records by participant identifier and time stamp. Descriptive data analysis was then used to provide an initial understanding of the participant's behavior over time and adherence to the intervention program. We analyzed the relationship between these different data sources and drew conclusions on the relationship between engagement with the DTx application and study outcome. To analyze the connection between the lifestyle intervention and OSA severity, the population was split into two groups based on whether their AHI over the 12-week intervention period changed and their engagement in the DTx application were compared. An decreased OSA severity was defined as a negative difference between the AHI after the study period and the AHI after the study period.

Association Rule Mining

To analyze the co-occurrence of lifestyle interventions, we applied association rule mining. Association rule mining is a machine learning technique used to uncover relationships between items in a dataset based on their co-occurrence [19]. Association rule mining is an unsupervised machine learning method that is commonly used to identify patterns in which items customers buy together or which pages of a website users open together.

Table 3: Binning of Numerical Features

Feature	**Bins↑**	**Bins↓**
Sleep Duration (hours	>8 h	<6 h
Awakenings (count)	>5Awakenings	0 Awakenings
Sleep Onset Latency (minutes)	>60 min	<10 min
Awake (hours)	>1 h	-
Screen Time (hours)	>8 h	<2 h
Sleep Quality (Likert scale, 1 lowest and 5 highest)	≥4	≤1
Staying Awake in Bed (minutes)	>60 min	<15 min
Activity (steps)	>10k steps	<1k steps
Stress (Likert scale, 1 lowest and 5 highest)	≥4	≤1

Frequently co-occurring items, known as frequent item sets, are used to identify rules with an "if-then" structure, e.g., if item X occurs, then item Y is likely to occur. In such rules, X is referred to as the antecedent, and Y is the consequent. These rules rely on three key metrics: *support*, *confidence*, and *lift*. Support measures the frequency of item co-occurrence, while confidence represents the likelihood of the consequent given the antecedent. Lift is a measure that tests the strength of an association rule by comparing its likelihood to random chance. The most common algorithm for generating association rules is the apriori algorithm [2]. It identifies frequent item sets by calculating support for all item combinations and retaining those meeting a minimum support threshold. These sets are expanded iteratively until no new frequent item sets are found. Rules are then derived by calculating confidence for each potential rule, filtering those that meet a minimum confidence threshold, and using lift to eliminate coincidental associations. Since association rules work with categorical data, all numerical

columns had to be binned. The binning was manually decided depending on the distribution of each feature. Table 3 shows how the sleep parameters were binned.

The data set was transformed into a transaction-based format, where each transaction referred to one night of sleep by one individual. The items in the transaction were the categorical features derived from the smartwatch and the DTx application. Merging the data from the smartwatch, the digital sleep diary and the DTx application, led to a transaction-based data set with 3642 transactions. In the following an exemplary transaction of Participant *n* on day *m* is showing a decreased sleep duration, a completed education Mission ($M_{\text{Education}}$) increased awakenings:

$$X_n Y_m = \{\text{Duration}\downarrow,\ M_{\text{Education}},\ \text{Awakenigs}\uparrow\}$$

The uniqueness of association rules is analyzing the data in the form of transactions and, therefore, handling complex data sets, including multiple features, of multiple participants on multiple nights. This allows us to analyze sleep, behavior, and engagement on a day-by-day level in this study. Given that participants could choose different missions based on their individual preferences, association rule mining enables the identification of frequent combinations of interventions used and related changes in sleep and behavior.

Results

In the following section, we will explain in detail how and how often participants engaged with the DTx application. Dividing the data by the changes in AHI further showed differences in the engagement with the application. Finally, based on the different intervention types, frequent patterns and association rules were derived.

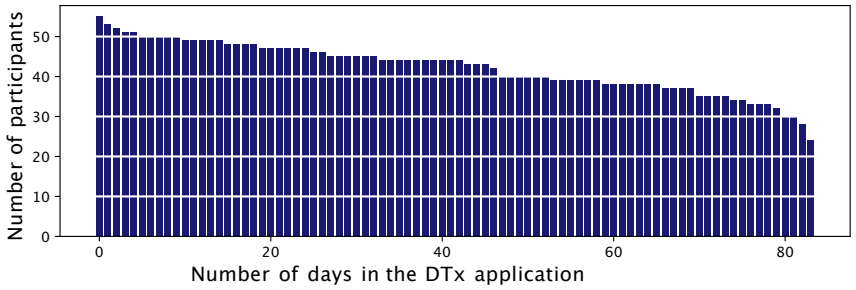


Figure 2: Participant retention in the app throughout the 12-week study duration Engagement with the DTx application

The first step in gaining insights into the effectiveness of DTx for improving sleep quality was collecting information about the 12-week study period. We analysed participant retention by identifying how many individuals completed the study, along with their levels of engagement with the DTx application throughout the intervention period. This included an analysis of the

frequency and consistency with which participants interacted with the app. Even though there was a dropout of participants over time, 43 participants adhered to the study design for more than half of the study period. Furthermore, 30 of the 51 participants used the application for more than 80 days. Figure 2 shows the drop out of participants over time.

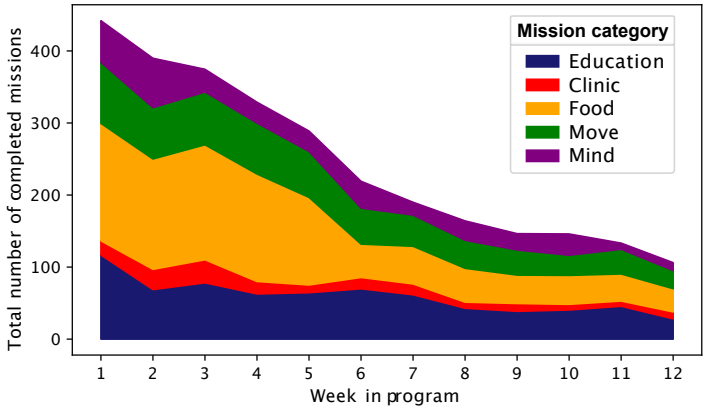


Figure 3: Average number of completed missions per week throughout the 12-week program, categorized by mission type

The data showed that the participants completed, on average, 375 missions in the duration of 12 weeks. This equals eight missions on average per day. However, the engagement with the DTx application decreased throughout the period of 12 weeks. Figure 3 shows how the average engagement of all participants decreases with each day of the program. The different colors show that the engagement varies between different types of missions. It is visible that the food missions were used most in the beginning but decreased over time. The education missions remained stable over the entire study period.

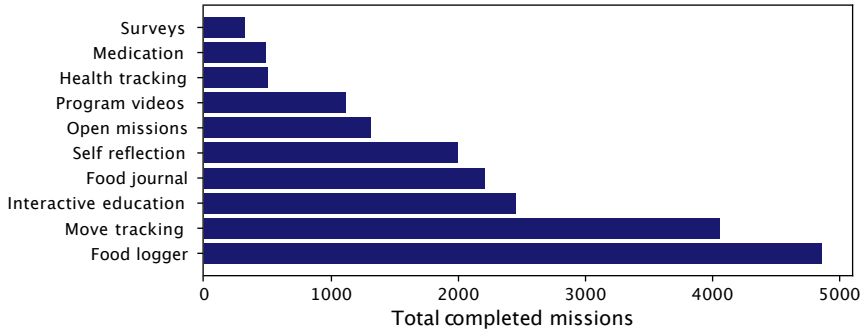


Figure 4: Total number of completed missions by mission type across the 12-week program

We can further split down the mission categories into specific missions. For example, both the food logger and food journal belong to the food mission category. Figure 4 shows an overview of the most completed missions. The missions that were most often completed by the

participants were logging food intake, reaching a step count goal, and watching interactive educational material. Comparing missions from the same category shows that participants more often log their food instead of using the food journal and more often use the interactive educational content instead of the program videos.

In the next phase of the analysis, we divided the study population into two groups. The goal of this division was to analyze whether the behavior of the participants with a successful reduction of OSA severity was different from the participants with no improvement of the OSA severance. The participants with a reduced OSA severity showed overall more engagement with the DTx intervention. The improvement group fulfilled on average 459 missions, while the non-improvement group fulfilled on average 345 missions. Dividing the missions by their mission category showed further differences in the behavior of the groups. The improvement group was more engaged in every mission category and exceeded, particularly in the food-related missions as can be seen in Figure 5. The groups also varied on the average number of days they retained in the study. The improved group (n=14) retained for on average 84 days in the intervention program, whereas the non-improvement group retained on average for 60 days in the program.

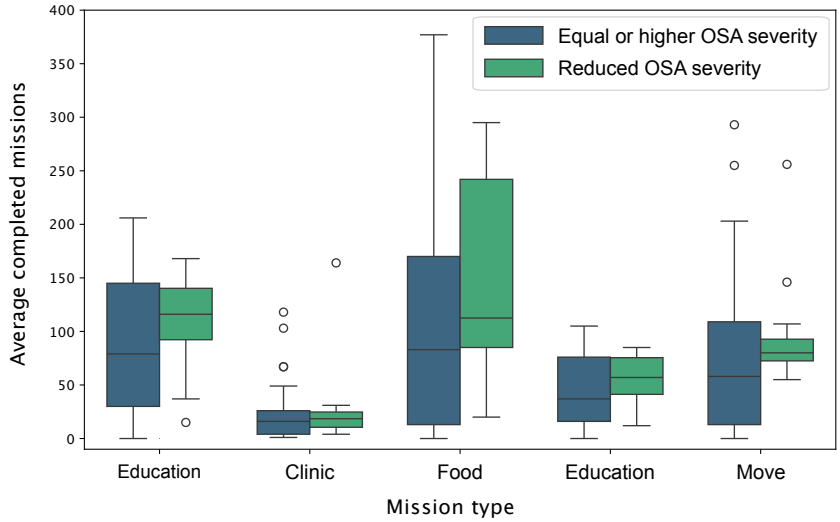


Figure 5: Engagement with missions by reduced OSA severity

Frequent Item Sets and Association Rules

Generating frequent item sets from the data set led to 81 sets, which have a minimum support of 0.1 and contain at least two elements. Table 4 shows five different identified frequent item sets. The item sets show that awakenings, snoozing, sleep duration, sending coach messages and stress show co-occurrences with missions in the DTx application. Particularly low awakenings, which refer to no awakenings during the night, frequently co-occur with the education and move missions. Also lowered snoozing, increased sleep duration and sending coach messages frequently co-occurred with the education missions. Increased stress frequently co-occurred with the move missions.

Table 5 shows examples of the association rules derived from the frequent item sets. In total, the apriori algorithm generated 398 rules with a minimum confidence of 0.1. Many rules appeared multiple times, switching the position of the items from antecedent to consequent and vice versa. Therefore, no strong direction of the association was found. The rules indicate that different missions co-occur with changes in sleep and behavior. These rules show, for example, that it is likely that a participant who completed a move mission and stayed only shortly in bed after they woke up is also more likely to complete an education mission and experience fewer awakenings during the night.

Table 4: Identified Frequent Item Sets

Frequent Item Set	Support	Number of Items
[Awakenings↓, $M_{Education}$, M_{Move}]	0.23	3
[Awake in Bed↓, $M_{Education}$]	0.22	2
[Sleep Duration↑, $M_{Education}$]	0.21	2
[Coach Message, $M_{Education}$]	0.15	2
[Stress↑, M_{Move}]	0.12	2

Another rule shows that participants were likely to engage with an education mission if they experienced improved sleep quality and engaged in a move mission. An interesting rule is that if a participant completes an education and move mission, it is more likely that this participant will also complete a food mission and experience higher sleep quality.

Table 5: Identified Association Rules

Antecedent	Consequent	Support	Conf.	Lift
[Awakenings↓, $M_{Education}$]	[Awake in Bed↓, M_{Move}]	0.1	0.40	1.83
[$M_{Education}$, M_{Move}]	[M_{Food}, Sleep Quality↑]	0.12	0.23	1.68
[$M_{Education}$, Screen Time↑]	[M_{Food}, M_{Move}]	0.1	0.79	1.75
[Pain↑, M_{Move}]	[$M_{Education}$]	0.11	0.95	1.55
[Sleep Duration↑, M_{Move}, M_{Food}]	[$M_{Education}$]	0.15	0.94	1.54
[Stress↑, M_{Move}]	[$M_{Education}$]	0.11	0.95	1.54

Discussion

The study showed that DTx can be an effective treatment approach for some but not all participants. However, looking only at the participants who decreased their OSA severity over the study period, an active engagement with the DTx application becomes visible. Therefore, we cannot conclude that this treatment path is a one-fits-all solution, but it can be a pathway for users who will independently engage themselves with the content and missions in the app. These results align with existing research suggesting that engagement plays a critical role for the success of DTx applications [5].

A key finding of this paper is the value of combining multiple data sources to create a comprehensive picture of the study. By being able to assess both the AHI as a clinical marker of OSA severity, in combination with the continuous tracking of both objective and subjective data, the study enriches substantial medical findings with longitudinal information on the

process the participant is undergoing in the 12-week period. The uniqueness of this study became clear when analyzing the engagement data with the DTx application. This data stream provides key information on whether the participant actively follows the lifestyle intervention and even more granular, how the participant changes their lifestyle.

The frequent item sets and association rules identified completed missions, sleep parameters, and user behavior that co-occurred frequently. An interesting rule is that completed move and education missions will likely co-occur with a completed food mission and improved sleep quality. It can be interpreted as that i) completing already two missions often goes together with completing another mission, and ii) completing multiple missions often goes together with improved sleep quality. This could show that completing different mission categories, i.e., working on different areas of pursuing a healthier lifestyle, could be a factor impacting sleep improvement. It is also an interesting finding that both the increased pain and increased stress co-occur with the move mission and are likely to go together with an education mission. These rules could show that participants who are in pain or stress and pursue activity or exercise are likely also to consume educational content.

There are some likely explanations for the observed decline in the average number of missions completed over time. First, it is common for users to engage more actively at the start of a program as they explore the app, familiarize themselves with its features, and discover which missions are most relevant to them. This initial engagement effect often decreases as users settle into a routine. Second, users tend to complete the missions they are served on a given day, and the number of available missions decreases over time. On average, users received 38 missions per week in the first three weeks, compared to 28 missions per week in the final three weeks. When multiplied by the number of participants, this structural difference alone can account for a significant decline in total mission completions. Third, users were also required to fill out a separate diary and use another app, which may have contributed to study fatigue as the program progressed. Finally, retention data shows a gradual drop in active users, which would naturally result in fewer missions being completed over time [20].

The observed differences in engagement raise broader questions about the factors driving user engagement. However, to truly understand why users select certain missions over others, it is essential to consider behavioral theories that explain intrinsic motivation. Integrating these theoretical insights with data-driven findings can inform the design of more effective DTx interventions for sustainable lifestyle changes. By applying association rule mining, the case study then provided insights into how digital therapeutics can be leveraged to promote healthier sleep habits. Association rule mining revealed that specific combinations, such as food-related and educational missions, often co-occur with changes in sleep or behavior.

Practical Implications

The implications derived from these results can be, on the one hand, useful for clinical practice and, on the other hand, useful for the designers and developers of these digital interventions. From a clinical perspective, this paper implied that DTx interventions can be a possible treatment path for OSA patients who will actively engage with the app. However, we cannot know beforehand who will engage and who will not. A potential research direction could be to connect the DTx application with offline coaching or physical exercises to also reach OSA patients who did not engage with the app. This way, these treatment paths could complement each other.

From a technical perspective, this paper implied that the combination of the DTx application together with continuous sleep tracking and self-reports from the digital diary, offers valuable

information to evaluate the application. The transaction-based is a useful way to structure the data, which is otherwise difficult to merge in a classical table format due to its three-dimensionality, considering participants, days and features. The identified rules imply further potential to apply this method to other longitudinal studies with multiple data sources. A possible future direction is sequential pattern mining, since the transactions have a temporal relationship, which is not considered in association rule mining.

Limitations and Future Research

One limitation of association rule mining is the simplification of numerical data into categorical data. In this step, some richness of the data is lost. Similarly, a generic binning of features neglects individual differences between the participants. For example, more than 8 hours of sleep may be a high sleep duration for many individuals but may be a low or average sleep duration for others. Therefore, applying personalized binning by including the distribution of values within individuals would be a good research direction for future work. Another limitation of this work is the limited knowledge about the parts of the DTx application that are not quantifiable. Features such as storytelling, design, or gamification have not been assessed in the data extracted from the user interaction and may be a subject for future qualitative research on how the users experienced the application.

The clinical validity of association rules is a further limitation of the study. The association rules are able to identify co-occurrences but do not imply causation. Therefore, the underlying cause-effect relationship remains unclear. The metrics support, confidence and lift can evaluate the reliability of these rules within this particular data set, but there is little research on how to translate these insights into practice. Future work could support the clinical validity of these rules by testing them with more broadly known statistical methods. Ultimately, the reliability of smartwatches poses a limitation, as we cannot verify how accurate the data on e.g. sleep onset latency or sleep duration is.

Conclusion

The paper showed that DTx interventions have the potential to reduce OSA severity, particularly when participants actively engage with the application. The different missions within the application showed that food, movement, and education were the most actively used features of the application and showed co-occurrence with awakenings, being awake in bed in the morning, stress, and sleep duration. The paper proposes association rules on a longitudinal data set as a possible way to evaluate digital intervention studies and shows the importance of combining multiple data sources to gain a comprehensive understanding of the participants and their behavior.

Acknowledgements

We thank everyone involved in the setup of the sleep study. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 965417. The corresponding author is Luka Biedebach and the shared senior authors are Jose Miguel Saavedra Garcia and Anna Sigríður Islind.

References

1. Abeysinghe R, Cui L. Query-constraint-based mining of association rules for exploratory analysis of clinical datasets in the National Sleep Research Resource. BMC medical informatics and decision making. 2018 Jul;18:58.
2. Agrawal R, Imieliński T, Swami A. Fast discovery of association rules. Advances in knowledge discovery and data mining. 1996;12(1):307-328.
3. Aji M, Gordon C, Stratton E, et al. Framework for the design engineering and clinical implementation and evaluation of mHealth apps for sleep disturbance: systematic review. Journal of medical Internet research. 2021;23(2):e24607.
4. Arnardottir ES, Islind AS, Óskarsdóttir M, et al. The Sleep Revolution project: the concept and objectives. Journal of sleep research. 2022;31(4):e13630.
5. Auster-Gussman LA, Lockhart C, Graham SA, et al. Engagement in Digital Health App-based prevention programs is associated with weight loss among adults age 65+. Frontiers in digital health. 2022;4:886783.
6. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. The Lancet Respiratory Medicine. 2019;7(8):687-698.
7. Berry RB, Albertario CL, Harding SM, et al. AASM Manual for the Scoring of Sleep and Associated Events. Darien, IL, USA: American Academy of Sleep Medicine; 2018. Version 2.5.
8. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep. 2014;37(1):9-17.
9. Camacho M, Certal V, Abdullatif J, Zaghi S, Ruoff CM, Capasso R, Kushida CA. Myofunctional therapy to treat obstructive sleep apnea: a systematic review and meta-analysis. Sleep. 2015 May 1;38(5):669-75.
10. Concaro S, Sacchi L, Cerra C, et al. Mining healthcare data with temporal association rules: Improvements and assessment for practical use. Artificial Intelligence in Medicine. 2009;12:16-25.
11. Dang A, Arora D, Rane P. Role of digital therapeutics and the changing future of healthcare. Journal of Family Medicine and Primary Care. 2020;9(5):2207-2213.
12. Espie CA, Kyle SD, Williams C, et al. A randomized, placebo-controlled trial of online cognitive behavioral therapy for chronic insomnia disorder delivered via an automated media-rich web application. Sleep. 2012;35(6):769-781.
13. Fridgeirsdottir KY, Bjórnsdóttir E, Ingadottir TS, et al. Effects of exercise and a lifestyle app on sleep-disordered breathing, physical health, and quality of life. ERJ Open Research. 2024.
14. López-Padrós C, Salord N, Alves C, Vilarrasa N, Gasa M, Planas R, Montsserrat M, Virgili MN, Rodríguez C, Pérez-Ramos S, López-Cadena E. Effectiveness of an intensive weight-loss program for severe OSA in patients undergoing CPAP treatment: a randomized controlled trial. Journal of Clinical Sleep Medicine. 2020 Apr 15;16(4):503-14.
15. Fürstenau D, Gersch M, Schreiter S. Digital therapeutics (DTx). Business & Information Systems Engineering. 2023;65(3):349-360.
16. Hilmarsdóttir E, Sigurðardóttir ÁK, Arnardóttir RH. A digital lifestyle program in outpatient treatment of type 2 diabetes: a randomized controlled study. Journal of diabetes science and technology. 2021;15(5):1134-1141.

17. Islind AS, Lundh Snis U, Lindroth T, et al. Individualized blended care for patients with colorectal cancer: the patient's view on informational support. Supportive Care in Cancer. 2021;29:3061-3067.
18. Kim JC, Chung K. Mining Based Time-Series Sleeping Pattern Analysis for Life Big-Data. Wireless Personal Communications. 2019;105(2):475-489.
19. Kotsiantis S, Kanellopoulos D. Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering. 2006;32(1):71-82.
20. Kristbergsdottir H, Schmitz L, Arnardottir ES, Islind AS. Evaluating User Compliance in Mobile Health Apps: Insights from a 90-Day Study Using a Digital Sleep Diary. Diagnostics. 2023 Sep 8;13(18):2883.
21. Laxminarayan P, Alvarez SA, Ruiz C, et al. Mining associations over human sleep time series. Proceedings - IEEE Symposium on Computer-Based Medical Systems. 2005;323-325.
22. Liang Z, Martell MAC, Nishimura T. Mining hidden correlations between sleep and lifestyle factors from quantified-self data. UbiComp 2016 Adjunct Proceedings. 2016;547-552.
23. Luik AI, Kyle SD, Espie CA. Digital cognitive behavioral therapy (dCBT) for insomnia: a state-of-the-science review. Current sleep medicine reports. 2017;3:48-56.
24. Makin S. The emerging world of digital therapeutics. Nature. 2019;573(7775):S106-S106.
25. Mclaughlin M, Delaney T, Hall A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. Journal of medical Internet research. 2021;23(2):e23180.
26. Murawski B, Wade L, Plotnikoff RC, et al. A systematic review and meta-analysis of cognitive and behavioral interventions to improve sleep health in adults without sleep disorders. Sleep medicine reviews. 2018;40:160-169.
27. Peppard PE, Young T, Barnet JH, et al. Increased prevalence of sleep-disordered breathing in adults. American journal of epidemiology. 2013;177(9):1006-1014.
28. Pevernagie DA, Gnidovec-Strazisar B, Grote L, et al. On the rise and fall of the apnea-hypopnea index: A historical review and critical appraisal. Journal of sleep research. 2020;29(4):e13066.
29. Prochnow L, Zimmermann S, Penzel T. Predictors of obstructive sleep apnea. Somnologie. 2016;20(2):113-118.
30. Punjabi NM. The epidemiology of adult obstructive sleep apnea. Proceedings of the American Thoracic Society. 2008;5(2):136-143.
31. Rusanen M, et al. Self-applied somnography: Technical feasibility of electroencephalography and electro-oculography signal characteristics in sleep staging of suspected sleep-disordered adults. Journal of Sleep Research. 2024;33(2):e13977.
32. Schmitz L, et al. Towards a Digital Sleep Diary Standard. Proceedings of the Americas Conference on Information Systems (AMCIS). Minneappolis; 2022 Aug 9-13.
33. Strollo PJ Jr, Rogers RM. Obstructive sleep apnea. New England Journal of Medicine. 1996;334(2):99-104.
34. Suzuki E, et al. Evaluation of an internet-based self-help program for better quality of sleep among Japanese workers: a randomized controlled trial. Journal of occupational health. 2008;50(5):387-399.

35. Thorgeirsson T, et al. Randomized trial for weight loss using a digital therapeutic application. Journal of diabetes science and technology. 2022;16(5):1150-1158.
36. Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. NPJ digital medicine. 2023;6(1):38.
37. Wickwire EM, et al. Patient Engagement and Provider Effectiveness of a Novel Sleep Telehealth Platform and Remote Monitoring Assessment in the US Military: Pilot Study Providing Evidence-Based Sleep Treatment Recommendations. JMIR formative research. 2023;7(1):e47356.
38. Willermark S, Islind AS. Choice Architecture, Friend, or Foe? Future Designers' Perspective on the Ethics of Digital Nudges. 13th Scandinavian Conference on Information Systems. 2022.