University of Southern Denmark

SDU

Accelerometry- and Temperature-Based Algorithms to Assess Sleep Habits Among Danish Children and Adolescents

PhD Thesis – October 2023

Esben Høegholm Lykke

Research Unit for Exercise Epidemiology, Centre of Research in Childhood Health, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark

Supervisor

Associate Professor

Jan Christian Brønd

Research Unit for Exercise Epedimiology, Centre of Research in Childhood Health, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Denmark

Assessment Committee

Chair

Professor WSR

Jasper Schiperijn

Research Unit of Active Living, Danish centre for motivation and behaviour science, Playground Research, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Denmark

Opponents

Associate Professor

Samuel Emil Schmidt

Department of Health Science and Technology, Aalborg University, Denmark

Associate Professor

Alex Rowlands

Biomedical Research Centre,
University of Leicester, United Kingdom

Funding

The research presented in this thesis was generously funded by TrygFonden, under grant numbers ID 130081 and 115606, and by the European Research Council, under grant number 716657. Additional support was provided by a one-year scholarship from the Faculty of Health Sciences, University of Southern Denmark.

Preface

The present thesis delves into the objective measurements of physical behavior and sleep, a subject that has captivated me since my Master's program. This work represents a fulfilling journey marked by exploration, discovery, struggles, and both personal and professional growth.

The thesis is based on data from several different studies and is the product of collaboration with numerous internal and external co-authors. It employs machine learning and advanced statistical methods on accelerometer data to bring new insights into the field of sleep research.

This thesis comprises three distinct papers, each focusing on improving and validating methods for leveraging accelerometer data in the study of human behaviors—particularly in relation to sleep. Each paper applies innovative methods, such as machine learning techniques, to enhance the utility, reliability, and accuracy of free-living accelerometer data in large-scale studies. Two of these papers have been published in peer-reviewed journals, and the third is under preparation. These works are integral to this thesis and are included as appendices.

My research journey began during my Master's program, where I was introduced to the capabilities of accelerometer data. This early exposure culminated in the publication of my Master's thesis and solidified my desire to pursue a career in research. Embarking on my PhD, I faced a series of challenges. Initially, my limited experience with programming and machine learning posed a steep learning curve. However, persistent effort enabled me to acquire the necessary skills for this kind of work. A significant hurdle arose during the data collection phase which was intended to be used in Paper III. I attempted to collect overnight polysomnography data, along with readings from multiple accelerometers and wrist photoplethysmography, from children in their homes. Regrettably, the sensitive nature of EEG electrodes did not mix well with children, resulting in data that was largely unsuitable for model development. After gathering data from 55 children and assessing its quality, we made the difficult decision to discontinue this line of data collection. Fortunately, I could turn to the SCREENS trial for alternative data, allowing me to complete the third paper.

The PhD experience has fundamentally shaped my approach to work and life, instilling in me qualities like discipline, precision, and a keen attention to detail. This journey has been as much about professional development as it has been a personal voyage of self-discovery and growth.

As I stand on the threshold of new beginnings, I am filled with excitement about the future possibilities. This thesis reflects the lessons and experiences gathered along the way and serves as a stepping stone for further exploration in this rapidly evolving field.

Enjoy reading.

Acknowledgments

As I reflect on the transformative journey that my PhD has been, I find myself indebted to numerous individuals whose support, guidance, and inspiration have been instrumental in shaping both my professional and personal growth. First and foremost, I extend my deepest gratitude to my Main Supervisor, Jan Christian Brønd. Your unwavering guidance and patience have not only nurtured my development as a researcher but also giving me the opportunity to lecture on interesting courses. Our collaborative dialogues, whether they took place in the office or during examinations, have been a cornerstone in my academic development.

To my co-supervisors, Anders Grøntved and Niels Christian Møller, as well as all the co-authors I've collaborated with: your valuable insights and unique perspectives have significantly enriched my work. Despite much of our collaboration being remote, the collegial spirit you displayed bridged the physical distance. I appreciate the camaraderie, mutual respect, and shared enthusiasm we maintained throughout our interactions, both in person and virtual. While I'm grateful for how seamlessly we adapted to remote work, I do wish circumstances would have allowed for more in-person interactions and shared moments. The occasions we spent time together emphasized how much I would have loved to hang out more. I hope the future brings more opportunities for us to collaborate and connect in person.

Being part of an internationally recognized and experienced research group has been an enriching experience. It afforded me the privilege to work alongside some of the most brilliant minds in my field. This collective experience has not only broadened my understanding but also contributed significantly to our shared goal of advancing knowledge in our discipline. My PhD journey has been about more than just academics. It's changed how I think and act in my everyday life. The careful, detailed way of doing research has affected how I make decisions and solve problems. This journey wasn't just about learning; it was also about understanding myself better. Every time I solved a tough problem, got my work published, or received good feedback, it reminded me why I'm doing this and how important it is.

In the heart of this journey lies the steadfast support of my family. To my wife, the cornerstone of our family, your steady support and interest in my work have been my emotional anchor. The joy and love from our four children have continually served as sources of inspiration and motivation.

I am incredibly grateful for all the support and wisdom I have been fortunate to receive, and it is my earnest hope that this thesis stands as a tribute to each of you. Thank you.

Included Papers

Paper I

Manual Annotation of Time in Bed Using Free-Living Recordings of Accelerometry Data

published in Sensors

Paper II

Generalizability and Performance of Methods to Detect Non–Wear with Free–Living Accelerometer Recordings

published in Scientific Reports

Paper III

Improving Sleep Quality Estimation: A Comparative Study of Machine Learning and Deep Learning Techniques Utilizing Free–Living Accelerometer Data from Thigh-Worn Devices and EEG-Based Sleep Tracking

In preparation for SLEEP

Table of contents

List of Figures

List of Tables

English Summary

Introduction: Sleep is an important element in promoting health, and the quantification of sleep has been improved with modern technology. Polysomnography, considered the gold standard, provides in-depth insight into sleep but is costly. In contrast, accelerometry is a cheaper and less invasive method, especially for longer home-based recordings. Machine learning is a tool that has the potential to automate and facilitate the estimation of sleep from accelerometer data. However, there are three challenges: producing reliable training data, ensuring data integrity through accurate removal of non-wear, and effectively using data to estimate sleep. Firstly, it is necessary to have sufficient and accurate annotations in the data for effective supervised machine learning, emphasizing the importance of methods for manual annotations based on accelerometer data. Secondly, it is essential to detect and remove periods when the device is not worn to perform accurate analyses. Identifying periods of non-wear is challenging, as traditional methods like logbooks can be prone to bias. Existing algorithms removes bias, but their accuracy is still debated. Finally, once data is correctly collected and processed, it is crucial to apply it effectively. Current methods for estimating sleep using accelerometers are based on data from wrist-worn and hip-worn devices, while data from thigh-worn accelerometers remains largely untapped for sleep estimation.

Objectives: Firstly, we will assess the accuracy of manual annotations for in-bed and out-of-bed timestamps in raw accelerometer data, comparing them to the timestamps determined by sleep diaries and an EEG-based sleep monitor. Secondly, we will assess heuristic algorithms and machine learning models for detecting non-wear. Finally, we will develop machine learning models for sleep classification and the estimation of sleep quality metrics using data from thigh-worn accelerometers and compare them with EEG-based sleep recordings.

Methods: For Paper I, accelerometer data from the hip and thigh of 14 children and 19 adults were used. Using Audacity, an open-source audio editing program, three raters annotated each accelerometer recording by marking the times when the person went to bed and when they got out of bed. Two rounds of annotations were performed to test reliability. The manual annotations were evaluated against both sleep diaries and EG-based sleep recordings. Concordance and agreement was evaluated using the intraclass correlation coefficient and Bland-Altman analyses.

Paper II used accelerometer data from sensors placed on the wrist, thigh, and hip. In hip and thigh data from 64 persons and wrist data from 42 participants, periods of non-wear were manually annotated in the same way as described in paper I. Three variants of decision trees were trained on 79.2% data from the hip and thigh and were evaluated against a selection of heuristic algorithms and recently developed machine learning models. The remaining data were used as test data for all included algorithms and models. Decision tree hyperparameters were optimized through five-fold cross-validation. External validation was performed on all wrist data. All included algorithms and models were evaluated using metrics derived from confusion matrices.

For Paper III, accelerometry and EEG-based sleep recordings from children aged 4-17 years were used. Data preprocessing included a low-pass Butterworth filter, removal of non-wear periods using the method described paper II, and a set of 64 predictors were extracted. Sleep recordings were median filtered in 5 and 10-minute windows before models were trained to better capture true awakenings. Two model strategies were used, a sequential approach with four pairs of binary classification models, and the other strategy used a multi-class model. Hyperparameter optimization was performed using ten-fold Monte Carlo cross-validation on the binary classifiers. Class imbalance was addressed using the synthetic minority oversampling technique. Data for training the multi-class model was split in a ratio of 50/25/25 for training, validation, and testing. For both strategies, the F1 score was used as an optimization target. Confusion matrix derivatives were used to assess epoch-to-epoch performance, and agreements on sleep quality metrics were assessed using Bland-Altman plots and Pearson correlations.

Results: The results of Paper I indicated excellent inter- and intra-rater agreement. Furthermore, the Bland–Altman limits of agreement were approximately ±30 min, showcasing only a minimal mean bias of manual annotation compared to EEG-based and sleep diary in-bed timestamps.

In Paper II, for detecting non-wear periods longer than 60 minutes, the established consecutive zeros algorithms were the most effective, registering F1-scores above 0.96. However, for durations shorter than 60 minutes, decision trees stood out, achieving F1-scores of over 0.74 across all sensor locations. Notably, the recently published deep learning and random forests models could not match this performance.

In Paper III, the XGBoost model performed the best when compared to an EEG-based sleep monitor in detecting sleep. The model demonstrated small biases in sleep period time (0.2 minutes), total sleep time (-7.0 minutes), sleep efficiency (-1.1%), and wake after sleep onset (-0.9 minutes). The model showed a moderate 0.66 correlation with total sleep time. Our limits of agreement for total sleep time, ranging from -95.5 to 81.4 minutes, were consistent with previous studies on hip and wrist devices.

Conclusions: Overall, the findings of this thesis underscore the reliability and precision of emerging technological methods in sleep and non-wear detection research. Paper 1 examined the agreement of manual annotations of in-bed time against traditional benchmarks and found it to be good to excellent across all comparisons. Paper 2 emphasized the nuances of non-wear detection, revealing clear strengths in certain algorithms for specific durations and highlighting areas where newer models need enhancement. Paper 3 highlights the XGBoost model for sleep assessment with thigh-worn accelerometers, situating it as a valid alternative compared to methods employed on hip and wrist accelerometer data. However,challenges remain in identifying in-bed awake periods and in assessing sleep quality metrics on an individual-basis, consistent with previous findings from wrist and hip-worn devices.

Dansk Resume

Introduktion: Søvn er et vigtigt element i sundhedsfremme og kvantificeringen af søvn er blevet forbedret med moderne teknologi. Polysomnografi betragtes som guldstandarden, og giver en dybdegående indsigt i søvn, men er omkostningsfuld. Omvendt er accelerometri en billigere og mindre invasiv metode, især til længere optagelser i hjemmet. Maskinlæring er et værktøj, der har potentialet til at automatisere og lette arbejdet med at estimere søvn fra accelerometridata. Dog er der tre udfordringer: at producere pålidelig træningsdata, sikre integriteten af data og effektivt bruge data til at estimere søvn. For det første er det nødvendigt at have tilstrækkeligt med nøjagtige annotationer i data for superviseret effektiv maskinlæring, hvilket understreger vigtigheden af metoder til manuelle annotationer baseret på accelerometridata. For det andet, for at udføre korrekte analyser, er det essentielt at detektere og fjerne perioder, hvor sensoren ikke er båret. Det kan være udfordrende at identificere perioder, hvor sensorene ikke bæres, da traditionelle metoder som logbøger kan være fejlbehæftede. Eksisterende algoritmer kan forbedre denne detektering, men deres nøjagtighed er stadig genstand for debat. Endelig, når data er blevet korrekt indsamlet og bearbejdet, er det afgørende at anvende det effektivt. Nuværende metoder til at estimere søvn ved brug accelerometre er baseret på data fra håndleds- og hoftebårne sensorer, mens data fra accelerometre, der bæres på låret, stort set er uudnyttede i forhold til at estimere søvn.

Formål: Denne afhandling har følgende formål. For det første vurderes præcisionen af manuel annotation af sengetider i accelerometridata sammenlignet med EEG-baserede sengetider og søvndagbøger. For det andet undersøges eksisternede og nye algoritmer og maskinlæringsmodeller til at detektere perioder, hvor accelerometeret ikke er båret. Endeligt udvikles maskinelæringsmodeller til søvnklassifikation og estimering af søvnkvalitetsmål ved brug af data fra accelerometre, der bæres på låret og sammenligner med EEG-baserede søvnoptagelser. Samlet set søger denne afhandling at forstå potentialet og udfordringerne ved at anvende maskinlæring til at estimere søvn via accelerometri.

Metoder: Til artikel I benyttedes accelerometerdata fra hofte og lår fra 14 børn og 19 voksne. Ved hjælp af Audacity, et open-source lydredigeringsprogram, annoterede tre bedømmere hver accelerometeroptagelserne ved at markere tidspunkter for, hvornår personen gik i sengen, og hvornår de stod ud af sengen. Der blev udført to runder med annotationer for at teste pålideligheden. 'Ground truth' baseredes på EEG-søvnoptagelserne. Overensstemmelse blev målt ved hjælp af intraklassekorrelationskoefficienten og Bland-Altman-analyser.

Artikel II anvendte accelerometerdata fra sensorer placeret på håndleddet, låret og hoften. På samme måde som beskrevet i artikel I, annoteredes perioder hvor sensorerne ikke blev båret i lår- og hoftedata fra 64 personer og håndledsdata fra 42 personer. Tre varianter af decision trees blev trænet på 79.2% data fra hofte og lår og det resterende data blev brugt til test. Hyperparametre blev optimeret gennem en fem-foldig krydsvalidering. Ekstern validering blev udført på al håndledsdata. Alle inkluderede algoritmer og modeller blev evalueret ved hjælp af mål afledt af confusion matricer.

Til artikel III benyttedes accelerometri og EEG-baserede søvnoptagelser fra børn i alderen 4-17 år. Dataforarbejdningen omfattede et lowpass Butterworth-filter, fjernelse af perioder, hvor sensorerne ikke blev båret via metode fra artikel II og et sæt på 64 prædiktorer blev konstrueret. Søvnoptagelserne blev medianfiltreret i 5 og 10 minutters vinduer inden modellerne blev trænet, for at fange sande opvågninger bedre. To model-strategier blev anvendt, en sekventiel tilgang med fire par af binære klassifikationsmodeller og den anden strategi anvendte en multiklasse model. Hyperparameteroptimering blev udført ved hjælp af ti-fold Monte Carlo krydsvalidering på de binære klssifikationsmodeller. En ubalance i træningsdata blev afhjulpet ved hjælp af synthetic minority oversampling technique. Data til træning af multiklasse-modellen blev opdelt i et forhold på 50/25/25 for træning, validering og test. For begge strategier blev F1 score anvendt som optimeringsmål. For at vurdere præstationen på alle modeller blev der anvendt mål afledt af confusion matricer og for at forstå effektiviteten af vores modeller til at estimere søvnkvalitetsmål blev Bland-Altman-plots og Pearson-korrelationer anvendt.

Resultater: Resultaterne fra artikel I viste fremragende enighed både mellem bedømmere og inden for samme bedømmer. Derudover var Bland-Altman limits of agreement cirka ±30 minutter samtidig med en minimal gennemsnitsbias for manuelle annotationer sammenlignet med EEG-baserede tidspunkter for tid i sengen og søvndøgbøger.

I artikel II, for perioder længere end 60 minutter var de etablerede algoritmer, som på forskellig vis detekterer perioder uden acceleration, de mest effektive og opnåede F1-score over 0,96. Decision trees viste sig at præstere bedst på perioder kortere end 60 minutter og opnåede en F1-score på over 0,74 på tværs af alle sensorplaceringer. De nyligt udviklede deep learning- og random forest-modeller kunne ikke matche disse resultater.

I artikel III vist XGBoost-modellen sig bedst til at bestemme søvn sammenlignet med EEG søvnoptagelser. Modellen viste små afvigelser i tid i sengen (0,2 minutter), total sovetid (-7,0 minutter), søvneffektivitet (-1,1%) og vågen efter først søvn (-0,9 minutter). Derudover viste denne model en moderat korrelation på 0,66 med total søvntid. Vores resultater vist limits of agreements som var sammenlignelige med tidligere studier på hofte- og håndledssensorer. Specifikt viste total søvntid limits of agreements på -95,5 minutter til 81,4 minutter.

Konklusion: Samlet set undersøger denne afhandling pålideligheden og præcisionen af metoder inden for bearbejdning af accelerometerdata og søvndetektering. Artikel I understreger at manuel annotatering stemmer overens med EEG-baserede og selvrapporterede sengetider. Artikel II fremhævede nuancerne ved detektering af perioder, hvor sensorerne ikke bæres og viste at visse metoder præsterer bedst for specifikke varigheder af perioderne. Artikel III fremhæver XGBoost-modellen som bedst til at klassificere søvn på data fra accelerometre på låret og viser sammenligelige resultater i forhold til metoder, der anvender maskinlæringsmodeller på data fra hofter og håndled. Dog er der stadig udfordringer med at identificere perioder, hvor man er vågen i sengen. Derudover understreger limits of agreements udfordringer i forhold til at vurdere individuelle søvnkvalitetsmål, hvilket er i tråd med tidligere fund fra sensorer, der bæres på håndleddet og hoften.

Introduction

Why Track Sleep in Health Research

Physical behaviors throughout a day encompass activities such as sleep, physical activity, and sedentary behavior. A plethora of research has emphasized the health benefits of optimal sleep, high levels of physical activity, minimal sedentary periods, and adequate sleep across all age groups[1–5]. These insights have informed public health guidelines on physical activity[6–8] and sleep duration[9–11].

Despite spending approximately a third of our lives asleep, many facets of sleep remain elusive[12]. What is clear is sleep's vital role in maintaining physical and psychological well-being, consolidating memories, and regulating emotions[13–15]. In contrast, insufficient sleep is associated with numerous negative health outcomes, from weight gain and heart disease to impaired immunity and elevated mortality risk[16–18]. Short-term consequences of poor sleep encompass reduced alertness, heightened stress, diminished concentration, and risk-taking behavior[19–22]. Chronic sleep deprivation can drastically reduce one's quality of life, increase accident risks, and have broader socio-economic repercussions[23–25]. Such concerns are intensified considering daytime sleepiness affects 10-20% of the population[26], attributed to factors like irregular sleep patterns, shift work, certain medical conditions, and medications[25].

Advancements in wearable technology now provide tools for in-depth insights into the intricate relationship between sleep, physical activity, and health[27]. Given the well-established importance of sleep and physical activity in overall well-being, integrating sleep tracking in broader health research is paramount. Modern wearable devices empower us to deeply understand the dynamics between sleep, physical activity, and overall health. Thus, comprehensive sleep research is crucial for a holistic understanding of a healthy life.

Determinants of Sleep in Relation to Health

Sleep is multifaceted, defined by its structure, duration, quality, and timing throughout the day. Human sleep consists of two primary phases: non-REM (Rapid Eye Movement) and REM sleep. Typically, healthy adults begin their sleep cycle in the non-REM phase, which itself is subdivided into three stages. As one transitions from stage to stage, the depth of sleep intensifies[28]. The third stage of non-REM sleep, often referred to as slow-wave or deep sleep, is particularly restorative and primarily occurs early in the night. Contrarily, REM sleep is characterized by increased brain activity and becomes more prolonged as the night advances[28]. Throughout the night, these sleep stages cycle, generally rotating between four to six times in intervals of roughly 90-120 minutes[28].

Disturbances in these sleep stages, such as interruptions from alarms, can have adverse health implications. Specifically, obstructing slow-wave sleep—even without changing

the overall sleep duration—can lead to reduced insulin sensitivity, poor glucose tolerance, increased sympathetic activity, and elevated morning cortisol levels[29,30].

The intricate relationship between sleep duration and health is evident in numerous epidemiological and experimental studies. Cross-sectional research has unveiled a "U"-shaped association where both shorter (typically less than 6 hours) and extended sleep durations (more than 8 hours) are linked with increased risks of obesity, mental health issues, coronary heart disease, stroke, and diabetes[4,31–34]. Furthermore, controlled experiments with sleep-deprived healthy adults have shown detrimental effects on their endocrine functions, leading to unfavorable metabolic and inflammatory responses[35,36].

Besides its duration, sleep quality is an essential component to consider. Factors determining quality include sleep onset latency, often known as latency until persistent sleep (SOL or LPS); wake after sleep onset (WASO); sleep efficiency (SE); and nocturnal awakenings[37]. Subpar sleep quality is associated with elevated risks of chronic diseases in adults, encompassing conditions like obesity, diabetes, and cardiovascular disease[38].

The growing body of research underscores the importance of understanding sleep's complexities and its pivotal role in health, reinforcing the importance of studying both its quantity and quality.

The Gold Standard for Measuring Sleep

The challenge of studying sleep has become significantly more manageable due to advancements in technology and our understanding of neuroscience. It wasn't until the 1950s that sleep study became scientifically feasible, thanks to the pioneering work of Nathaniel Kleitman and Eugene Aserinsky[39], who demonstrated the brain's active involvement in sleep. Utilizing electroencephalography (EEG), Kleitman and Aserinsky were able to measure brain activity and identified that it synchronizes over multiple regions, predominantly within specific frequency bands. This groundbreaking discovery enabled them to define distinct "sleep stages" that the brain cycles through, fundamentally transforming the way sleep is measured and understood. A visual example of these sleep stage cycles can be seen in Figure 1. Therefore, measuring sleep, given its complexities, not only is critical but also possible thanks to these technological and scientific milestones.

In a relaxed wakeful state, the EEG predominantly displays alpha activity within the frequency band of 8-10 Hz and amplitudes usually ranging from 10-50 μV. As an individual begins to fall asleep, they enter the non-rapid eye movement (NREM) sleep stage 1 (N1). During this drowsy phase, the brain's EEG activity transitions to the theta range frequencies of 4-6 Hz. Simultaneously, muscle relaxation is evident, respiratory rates decelerate, and there's a drop in both distal body temperature and heart rate. This initial stage is followed by NREM sleep stage 2 (N2), where the EEG spectra frequencies are further reduced. This stage introduces sleep spindles—periodic high-frequency waves oscillating between 12-14 Hz lasting for 0.5-1.5 seconds—and K-complexes, which are characteristic reactive EEG elements of N2 sleep.

Progressing deeper into sleep, one enters NREM sleep stage 3 (N3), often termed "deep sleep". Here, the EEG mainly exhibits high-amplitude oscillations within the delta band

of 1-4 Hz. Following these NREM stages, the sleep cycle culminates in REM sleep. Interestingly, the EEG patterns during REM sleep closely resemble the alpha activity seen in wakeful states. This stage is marked by evident rhythmic eye movements, while the respiration and heart rate display enhanced amplitudes and variability. The brain stem suppresses most voluntary body movements at this time, making it a period of relative physical inactivity. Dreaming tends to be more frequent during REM sleep. A single REM episode can span a few minutes and tends to extend in duration during the latter part of the night. On average, an entire sleep cycle, from N1 to REM, spans 90-110 minutes and is typically repeated multiple times throughout the night.

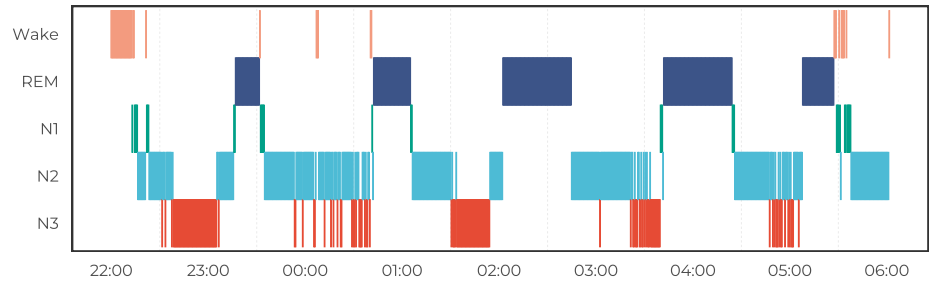

Figure 1: Sample hypnogram showing the sleep stage cycles of an eight-hour polysomnography recording. The sleep stages (REM, NREM 1-3) and arousals are shown.

Polysomnography (PSG) is a gold-standard technique in sleep research, allowing simultaneous assessment of various physiological signals influenced during sleep[40]. PSG gathers electrophysiological data from the brain using a 6-channel EEG, specifically from locations F3, F4, C3, C4, O1, and O2, contrasted against the contralateral mastoids (M1, M2). In addition, it uses electrooculography (EOG) to assess eye movements, electromyography (EMG) to track chin muscle tone and occasional arm and leg movements, and electrocardiography (ECG) to monitor heart rate. The study is augmented by methods to assess respiratory airflow, respiratory effort indicators, as well as peripheral pulse oximetry (PPG)[41]. For enhanced data interpretation, an infrared-equipped video camera captures the sleeping subject. Typically, PSG is carried out overnight in a specialized clinical sleep laboratory, assisting in diagnosing various sleep disorders. This technique provides detailed insights into an individual's sleep architecture, revealing sleep and wake durations as well as aiding in the classification of sleep stages[40]. Such detailed data enables accurate clinical research and the diagnosis of various sleep disorders, such as sleep apnea and periodic movements during sleep[40]. While PSG offers an unparalleled depth of sleep data, essential for diagnosing an array of sleep disorders, it comes with its own set of limitations. The procedure can be costly, often restricted to one or two nights in a specialized setting under a technician's supervision. This controlled environment may not truly mirror free-living sleep conditions[40]. Moreover, PSG necessitates specialized personnel to oversee, score, and interpret the data, making it less feasible for expansive or free-living studies[42] and also introducing inter and intra-rater differences in scoring the PSG data[43]. Hence, while invaluable, PSG is predominantly reserved for individuals presenting sleep-related complaints and for the conclusive diagnosis of sleep disorders.

The Zmachine® Insight+

Automated EEG data scoring presents a cost-effective alternative that mitigates the subjectivity tied to manual scoring by technicians[44]. While there has been an uptick in technological advancements recently, substantial progress is still needed to create objective, dependable, and valid methods to determine sleep metrics[45]. From the relatively few studies on automated scoring algorithms, some have shown encouraging outcomes. For instance, Malhotra et al.[46] explored an automated system, comparing its PSG data scoring with visual evaluations done by PSG professionals. They found that their computer-based method yielded outcomes comparable to those of seasoned technologists. Yet, this algorithm relies on several physiological data channels, including EEG, chin EMG, and electrooculography. In the past decade, single-channel EEG-based sleep staging algorithms have started to gain attention among researchers who have proposed a variety of potential scoring methods that are compared against traditional visual scoring[47–49].

The Zmachine® Insight+ (ZM) emerges as an important asset in sleep studies. With positive validation against PSG[50,51], the ZM delivers data on par with this gold standard, but without the hefty expenses or the demand for specialized oversight typical of PSG. Notably, the ZM's user-friendliness[52] allows for multi-night evaluations in real-world settings, capturing genuine sleep pattern fluctuations. This offers an edge over one-night PSG studies, making it an ideal primary data source for machine learning analyses. This is because it offers several nights of data without inconsistencies from different raters. However, despite its advantages, the ZM, much like the PSG, still has considerable costs and demands on participants. This underscores the importance of more convenient and cost-effective options.

Sleep Questionnaires and Diaries

Sleep has traditionally been assessed using a sleep questionnaire in larger-scale studies. These are cost-effective and quick, making them suitable for first-line diagnosis and assessments. They quantify a patient's subjective perception of their sleep quality. While these questionnaires are inherently subjective, they've been validated as accurate in numerous studies[53–57]. Typically, medical professionals are not needed to administer these questionnaires; they can be self-completed, even at home. For example, several apps exist that instantly provides a report after questionnaire completion, assisting those with potential sleep issues to seek specialist care. It's vital to understand that not all questionnaires measure the same aspect of sleep. While some assess sleep quality, others like the FOSQ-10 evaluate sleepiness[58] whereas instruments like the Pittsburgh Sleep Quality Index offers insights into an individual's overall satisfaction with their sleep over a defined time-frame, often a month[40]. In population studies, self-report sleep assessments are common but flawed. They can overestimate sleep duration and miss subtle sleep quality details. Their design, summarizing sleep data over weeks, risks recall biases, especially when remembering older sleep patterns[40]. Factors like weight, ethnicity, and regular sleep duration can influence these self-reports' accuracy[59].

Stepping away from these broad self-reports, sleep diaries stand out as a more detailed and structured tool. Often framed as the "gold standard" in subjective sleep assessment, they dive deep into various sleep parameters, like total sleep duration, efficiency, onset latency, and wake periods post sleep onset. Their strength lies in offering a day-by-day account, making it easier to spot disturbances, ascertain precise sleep timings, and decipher the rhythm of daily sleep-wake patterns over an extended duration[41]. However, like all tools, they aren't perfect. Their accuracy hinges on participants' memory retention and commitment to regular, detailed diary entries. From the researchers' standpoint, sifting through these extensive diaries can be time-intensive and, for participants, the process can sometimes be seen as taxing, potentially affecting their consistency in logging entries[60].

Accelerometry for Assessing Sleep

To address the limitations of EEG-based systems and self-reported sleep assessments, accelerometers have emerged as a valuable alternative. Generally, an accelerometer is an electronic device that measures both static and dynamic acceleration forces. Static forces, arising from gravitational pull, let devices determine orientation, like a smartphone's landscape or portrait mode. Dynamic forces, resulting from movement or vibrations, are utilized in activities such as step-counting in fitness trackers or detecting collisions in vehicle airbag systems. Through technologies like piezoelectric and MEMS (micro-electromechanical systems), accelerometers measure acceleration across various axes, serving diverse applications from aerospace to consumer electronics. Essentially, they are crucial sensors relaying motion or position changes to electronic systems.

While sleep researchers refer to these as "actigraphy devices", those in physical activity studies call them "accelerometers." Both terms denote devices employing accelerometer sensors to detect motion. For physical activity measurement and physical activity type distinction, devices are often placed on the hip or thigh[61–65], mainly detecting vertical acceleration associated with walking or running. This stemmed from early studies using uni-axial accelerometers that sensed movement in one direction. Devices for sleep, however, are usually wrist-worn. Omnidirectional accelerometers can sense movement in multiple directions, giving a composite signal, whereas triaxial accelerometers, with three orthogonal units, measure acceleration in three planes[66]. These tools can offer objective insights into sleep patterns in free-living settings over consecutive nights[67]. Their affordability and non-intrusive nature make them more appealing than PSG systems for population studies. However, many studies focus exclusively on sleep or activity, leading participants to wear the device either during the day for activity or at night for sleep. This can result in inaccuracies, like not wearing the device immediately upon waking or removing it before sleep[68]. To improve consistency, guidelines recommend 24-hour device wear[69].

Over the last 30 years, several research studies have highlighted the dependability and accuracy of actigraphy as an alternative to PSG for determining nocturnal sleep-wake patterns[70–77]. The findings from these investigations indicate a consistent epoch-by-epoch concordance of 80 to 95% between accelerometer-based sleep-wake scoring methods and the conventional PSG-based scoring. Due to this high degree of accuracy, the

inclusion of actigraphy devices has become a standard in sleep medicine protocols for diagnosing various sleep disorders[78].

Actigraphy employs algorithms to distinguish sleep from wake states, using movement as an indicator of wakefulness. Though these algorithms vary depending on factors like device brand and placement, they share a common principle: they categorize each epoch based on surrounding activity levels. Early actigraphy devices, due to technical constraints, converted raw acceleration data into activity counts for storage[79]. The first of such algorithms emerged in 1982, validated against PSG[80]. Its successor, the Cole-Kripke algorithm, became widely accepted by 1992[71]. These algorithms typically analyzed activity count-based features around a specific time frame and utilized linear or logistic regression for binary sleep-wake predictions[76]. In response to technological evolution and research demands, manufacturers began offering raw acceleration data from actigraphy devices. This shift allowed the creation of sleep algorithms rooted directly in raw acceleration rather than aggregated activity counts.

Borazio et al. introduced the Estimation of Stationary Sleep-segments (ESS) algorithm, designed for raw accelerometry data. This algorithm identifies idle segments by detecting a sustained low standard deviation per second over a minimum of 10 minutes[81]. On a related note, van Hees et al.[82] developed an algorithm centered on the accelerometer's estimated orientation angle. It identifies time segments where the estimated angle relative to gravity remains within a 5 degrees variance for at least 5 minutes. This method offers a more intuitive understanding than traditional methods, which typically focus on acceleration magnitude and zero-crossing counts. This heuristic algorithm has since gained widespread adoption in research[83–86]. More presently an algorithm developed for data collected from thigh-worn devices that relies on a constantly changing variable called 'sleep index' that is affected by movement[87]. While heuristic methods have demonstrated their merit, they inherently don't benefit from increased data availability. With the growing adoption of accelerometer devices, we're seeing an influx of data. This suggests that a shift to machine learning approaches could be advantageous, as they offer advanced capabilities to leverage the full potential of these vast datasets.

Machine Learning Fundamentals

Machine learning, a subfield of artificial intelligence, revolves around the design and implementation of algorithms that empower computers to extract patterns from data and, consequently, make informed predictions or decisions[88]. These algorithms, distinct from conventional explicit programming, deploy statistical methods to discern patterns within data[89]. Such iterative and experiential learning allows computational systems to progressively enhance their performance, autonomously adapting to new insights and data.

Different methodologies fall under the broad umbrella of machine learning, which include supervised, unsupervised, and reinforcement learning. In supervised learning, the algorithm has access to both input data and its corresponding desired output. The model, in a "supervised" fashion, learns from this data pairing until it can adeptly predict outputs for previously unseen data. Unsupervised learning, in contrast, works with unlabeled data

and aims to identify hidden structures or patterns within this information. Then there's reinforcement learning, a paradigm wherein a model fine-tunes its strategies by interacting with an environment, gathering feedback in the form of rewards or penalties[90]. Within the confines of this thesis, our spotlight is on supervised learning, especially its applications in pinpointing behaviors via accelerometry data.

Central to the philosophy of machine learning is the principle of training. Equipped with a designated training dataset, the algorithm continually predicts and recalibrates its strategies based on any discrepancies or errors it encounters. Over time, this repetitive adjustment sharpens the algorithm's proficiency, whether it's tasked with recognizing images, decoding spoken language, predicting sales trends, or identifying sleep patterns in accelerometry data.

Deep learning represents a specialized branch of machine learning. It capitalizes on multi-layered neural networks, giving it the label "deep", to discern complex patterns within datasets. Two standout architectures in this domain are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs excel in identifying spatial structures and pinpoint patterns, whereas RNNs are tailored to capture time-based or sequential information. This level of adaptability and precision allows deep learning models to analyze a wide variety of data with exceptional accuracy and detail[91].

The machine learning landscape is dotted with a diverse array of algorithms tailored for a myriad of supervised learning challenges, spanning from elementary linear regressions to intricate deep learning models. Every algorithm has its niche strengths, but they collectively share the goal of learning patterns from data to yield accurate predictions and insights. For instance, Linear Regression is fundamental, forecasting a continuous result variable based on one or several predictor variables with an assumed linear relationship between them[92]. Support Vector Machines cater to classification and regression by determining the optimal hyperplane that best divides datasets into discrete classes[93]. On the unsupervised end of the spectrum, K-Means Clustering strives to categorize data into clusters grounded in similarity[94]. Although the world of machine learning boasts a seemingly infinite repository of learning algorithms, this thesis narrows its focus to select algorithms that underpin the models we've developed.

1. Logistic Regression: Contrary to its name, logistic regression is used for binary classification rather than regression. Logistic Regression estimates a linear decision boundary (or hyperplane in higher dimensions) to differentiate between classes, and it models the probability that a given instance belongs to a particular class using the logistic function applied to a linear combination of the input features.[95].

2. Decision Trees: Decision trees offer a graphical representation of possible outcomes to a decision, making them easily interpretable. By segmenting the dataset based on feature values, they can handle both categorical and numerical data. They are employed in diverse areas, including medical diagnosis and credit risk analysis[96].

3. Multilayer Perceptron: Often termed a single-layer, feed-forward neural network, multilayer perceptrons consist of at least three layers: an input layer, a hidden layer, and an output layer. They can model complex relationships by adjusting weights between nodes during training. Activation functions, like the sigmoid or ReLU, introduce non-linearity into the model[91].

4. XGBoost: XGBoost is an optimized gradient boosting machine learning algorithm known for its speed and performance. It operates by sequentially adding weak learners, typically decision trees, and making corrections based on errors from previous stages, thereby producing a robust prediction model. This algorithm has notably excelled in various Kaggle competitions and is particularly favored for analyzing structured or tabular data[97]

5. Bidirectional LSTM (Long Short-Term Memory): LSTM networks, a type of RNN, are designed to remember patterns over long sequences and are particularly effective for time-series and natural language processing tasks. The bidirectional variant processes the sequence data from both past-to-future and future-to-past directions, enhancing the context information available to the network[98].

Harnessing Data for Sleep Detection Insights

Machine learning have reshaped various sectors, ranging from healthcare and finance to entertainment and technology. Central to this transformative wave is data, which has been metaphorically dubbed the new "oil" of the digital age, a term introduced by Clive Humby in 2006.

The potency of machine learning models hinges on the quality, volume, and diversity of the training data. This dependence on data has a dual rationale. Firstly, by their nature, machine learning algorithms discern patterns and relationships within data. A dataset that's both rich and varied enhances the algorithm's ability to generalize and make precise predictions on unfamiliar data. However, a sparse dataset can lead to overfitting, causing the model to become too tailored to the training data and thus perform inadequately in real-world situations. Secondly, machine learning algorithms, especially deep learning models, which often employ multi-layered neural networks, have a vast number of parameters which demands a substantial volume of data for effective training. It's akin to fitting pixels in an image together—the more pixels (or data) you have, the clearer the overall picture (or pattern) becomes. Nevertheless, sheer volume isn't the sole metric of value. The data's quality, relevance, and range are of paramount importance. Comprehensive, high-caliber datasets, capturing a wide range of scenarios and edge cases, lay the foundation for robust, adaptive, and trustworthy machine learning models.

With the surge in data availability from wearable devices, the door has opened wide for integrating machine learning methodologies into accelerometry research. These sophisticated algorithms can sift through expansive and intricate datasets, revealing insights that were previously obscured. Machine learning offers enhanced precision and depth to data analysis, whether it's pinpointing nuanced patterns of sleep apnea for early intervention[4] or assessing the efficacy of antidepressants in those afflicted with sleep disturbances[10]. The sphere of sleep research has benefited from machine learning and deep learning techniques[73,77,99,100], particularly in the classification of sleep stages. Notably, Tilmanne et al.[73] demonstrated that Multilayer Perceptrons and Decision Trees outperformed the algorithms of Sazonov and Sadeh[70,76]. Similarly, Granovsky et al.[77] employed Convolutional Neural Networks to distinguish sleep-wake patterns, though their

results diverged from peer research due to the lack of PSG benchmarking. Additionally, deep learning models have showcased superior accuracy over conventional models in datasets like the MESA sleep[99,101]. The GGIR[102] R package originally incorporated the algorithm by van Hees et al. from 2015[82]. However, recent advancements have seen a random forest model, also now integrated into GGIR, surpassing the van Hees model in accuracy. This new model also outperforms other well-established algorithms such as Cole-Kripke and Sadeh[100]. Recently, Trevenen et al. applied machine learning for sleep classification, extracting numerous features from the acceleration vector magnitude. They utilized these features in a Hidden Markov Model to distinguish between sleep and wakefulness and to identify the four sleep stages[103]. While their innovative approach to solely using accelerometer data for sleep stage classification didn't yield high accuracy, especially in detecting REM and differentiating Non-REM stages, they remained hopeful about the potential of such classifications.

While progress has been made, there are still challenges in creating machine learning models for sleep detection. We'll discuss these challenges in the following sections.

Importance of Accurate Data Annotation

The complexity of a machine learning model is closely tied to the number of model parameters it must learn. As the number of features a model considers increases, there's a proportional demand for more data. For instance, in predicting housing prices where a model evaluates variables such as location, number of bedrooms, and the neighborhood, a diverse dataset is important to understand the influence of each variable[88]. While basic learning algorithms can often produce satisfactory outcomes with relatively limited data, more intricate learning algorithms, especially those within the deep learning spectrum, have a heightened data requirement. One of the standout attributes of deep learning, in contrast to traditional machine learning, is its ability to draw insights directly from raw data without the need for manual feature engineering. This capability necessitates a richer and more diverse dataset for optimal model performance[91]. Data volume also depends on the task's complexity and the acceptable error margins for the application. A weather prediction model might tolerate a 20% error, but medical diagnostics require near-perfect accuracy. Lastly, the unpredictability or diversity of input can significantly influence data needs. Take, for instance, an online virtual assistant. Given that users can pose a myriad of queries in various styles and with occasional grammatical errors, the underlying model must be trained on a broad dataset to handle this range of unpredictability[91].

The complexities and nuances associated with data requirements in machine learning underscore the importance of not only having the right volume and diversity of data but also ensuring its quality and precision[91]. One of the pivotal aspects ensuring this precision, particularly in supervised learning scenarios, is data labeling. It means marking data with specific tags, guiding models to learn and predict. These labels, which are typically manually added by experts, help the models identify patterns. They can indicate things like categories, feelings, or any task-specific information. The better the annotation, the better the model performs, so thorough annotations is essential. Yet, the manual

annotation process, especially by experts, can be complex. Simple visualization might not always offer a comprehensive understanding, potentially leading to inaccurate labels. With the vast data volumes at play, this procedure is not just lengthy but could also be error-prone, especially for lengthy annotation tasks. Inherent limitations in the precision of manual labeling mean that any missteps can profoundly skew results.

In the context of machine learning models tailored for sleep detection in multi-day accelerometry recordings, accurately annotating bedtimes and wake-up moments is essential. While one might consider sourcing annotations from sleep diaries or EEG recordings, many studies have not integrated these within their study design, leading to an information gap which can be circumvented via manually annotating targets of importance. Moreover, many current methods aimed at detecting sleep don't effectively capture the 'in-bed' time also termed the sleep period time. An exception is the HDCZA algorithm by Van Hees et al.[104]. The question then becomes whether manual annotations of 'in-bed' times in accelerometry data can serve as a reliable alternative to sleep diaries or EEG recordings. If they can, this would offer a means to enrich a vast amount of existing data that currently lacks associated sleep logs or other 'in-bed' time indicators, making it more suitable for machine learning tasks. However, as of now, there's no research showcasing the effectiveness and precision of such manual annotations.

Integrity of Accelerometry Data

Data integrity is fundamental to any credible research or analytic endeavor, a principle that's especially evident in accelerometry. Here, the accuracy and completeness of data shape our understanding of human motion and behavior. While it may seem basic, mounting the accelerometry devices correctly—using tape, elastic belts, or other secure mechanisms—is essential. Improper mounting can lead to errors, such as a device being flipped or wrongly reattached after a non-wear period, ultimately compromising the study's results and conclusions.

Non-wear time represents a 'missing data' challenge in accelerometry datasets. This issue arises when devices aren't worn due to activities like swimming, sleeping, or malfunctions, among other reasons. Notably, a valid data day is defined by having at least 10 hours of wear time, and participants must have data from at least one weekend day within a minimum of four valid days[61]. Addressing non-wear time is crucial in data processing, especially when ensuring the accurate differentiation between true non-wear periods and sleep[105,106]. Furthermore, as researchers derive secondary parameters for physical activity, precision in distinguishing between wear and non-wear times becomes paramount[107]. Misclassifications can lead to skewed activity estimates, affecting the reliability of the conclusions[108].

To handle the non-wear time challenge, some researchers have participants maintain a log diary, though this method has its drawbacks, including potential errors and participant burden[109]. In pursuit of accuracy, the community has explored both rule-based methods and advanced algorithms. An early strategy, specific to ActiGraph data, identified non-wear time by examining consecutive zero counts[110–112]. However, minor changes in thresholds can significantly alter outcomes[113]. The proprietary algorithms in this field

have also faced scrutiny due to transparency issues and variable influences like age and obesity, affecting cross-study comparability[114].

Technological advancements in accelerometers have enabled them to store raw acceleration data, promising more detailed data and refined non-wear classifications. Various algorithms have emerged to decode this raw data, with some integrating skin temperature for enhanced accuracy[115–117]. The open-source GGIR package stands out, offering both a means to detect non-wear time and a method to adress non-wear time by substituting it with imputed values derived from averages of similar time points from other measurement days[102]. Other statistical imputation techniques, like the zero-inflated Poisson and Log-normal distributions, provide alternative solutions but come with inherent biases[118].

While these heuristic methods show broad applicability across diverse datasets and devices, their primary challenge lies in potential misclassifications due to set time intervals. Furthermore, data quantity and quality don't always ensure improved outcomes—a contrast to machine learning models that benefit from increased data.

Emerging technologies have ushered in machine learning methods, like random forests[100] and algorithms involving convolutional neural networks[119], optimized for raw accelerometer data classification. However, these models must tread the fine line between variance and bias, ensuring they don't overfit or underfit. Despite their efficacy in testing, these models' generalizability to new, unseen data remains uncertain. This brings the essentiality of external validation to the forefront, a step often overshadowed due to data constraints or research design choices. Some studies combine skin temperature with raw data for improved classification[117], but the full effects of this combination in machine learning remain largely unexplored. The quest continues for the ideal algorithm or model, one that flawlessly classifies non-wear time in raw accelerometer data across diverse settings.

Limitations of Current ML Models to Detect Sleep

Although machine learning have been applied to accelerometer data with the goal of predicting sleep for a decade, the field is still in its infancy. Primarily, most methods have been tailored to integrate with data sourced from wrist- and hip-worn devices[67].

The use of machine learning and deep learning on data collected via devices mounted to the thigh remains unexplored despite the potential advantages of this sensor location in estimating physical activity behaviors[64]. Only a few studies have leveraged this sensor-location for heuristic algorithms[87,106,120–122]. Given that specific sleep-related behaviors or positions might be better captured by thigh-mounted devices compared to their wrist or hip counterparts, one must wonder how this skewed focus impacts the adaptability and performance of these models in diverse real-world scenarios. Not leveraging this potential data source might lead to overlooked nuances in sleep detection.

A significant hurdle in assessing sleep using accelerometer data is the extraction of the sleep period time window without supplementary data from sleep logs or diaries[123–125].

This reliance on subjective inputs introduces biases, assuming that participants consistently log accurate timings. This approach can be especially problematic in long-term studies or specific demographics, like children who often rely on their parents for accurate bedtime reporting. Typical methodologies that depend on accelerometers often require participants to log their bedtime, sleep initiation, and wake-up moments diligently[42,75,126], which can be burdensome and might result in incomplete or inaccurate datasets.

The foundational data for many algorithms, such as the random forests model by Sundarajan et al.[100], is often based on single-night PSG-recordings. While useful for a snapshot of sleep behavior, this approach doesn't account for night-to-night variability in sleep patterns. Encapsulating intra-individual variances across multiple nights could foster more robust and generalized models. However, repeated nights of PSG recording can be challenging due to its intrusiveness, cost, and inconvenience for participants. In this light, alternative systems, like the ZM, emerge as more practical and less intrusive than traditional PSG.

Given these existing challenges and opportunities in the field of accelerometry for sleep detection, it becomes evident that there's an scope for innovation and improvement. By delving into these areas of potential, researchers can foster a holistic understanding of sleep, sedentary behavior, and physical activity. With this backdrop of both the constraints and the immense possibilities that the field offers, we now turn our focus to the central ambitions of this thesis.

Thesis Aim and Objectives

The rapidly growing field of wearable technology provides opportunities to collect accurate and objective data on human behavior, particularly using free-living accelerometer recordings. This thesis situates itself within this developing domain, with the ambition of harnessing the potential of wearable accelerometer technology for sleep estimation. At its core, the overall aim is to innovate methods and models for the analysis and interpretation of sleep. Building on the challenges delineated in earlier sections, this thesis endeavors to advance the field through a series of papers, as detailed below.

- The objectives of paper I is to present a method for manually annotating individual bedtime and wake-up times using raw accelerometry data. Furthermore, to validate the accuracy of these annotations by comparing them with in-bed and out-of-bed timestamps as determined by the ZM and by a prospective sleep diary, and lastly, to assess both inter-rater and intra-rater agreement of the manual annotations.
- Paper II lays down two primary objectives. Firstly, it seeks to evaluate decision tree models developed from data from both thigh and hip-worn accelerometers to detect non-wear time in accelerometry data, also including the role of surface skin temperature. Secondly, it draws a comparison between machine-learned models and heuristic algorithms across accelerometer datasets sourced from devices worn on both the hip, thigh, and wrist.
- In Paper III, the primary objective is to assess the performance of a selection of machine learning and deep learning models in estimating in-bed and sleep time, benchmarking all included methods against sleep recordings of the ZM. Beyond this, the secondary objective is to evaluate the ability of the developed models to quantify commonly used sleep quality metrics, once again validated against an EEG-based sleep tracking device.

Paper I: Manual Annotation of Time in Bed Using Free-Living Recordings of Accelerometry Data

This segment of the thesis encompasses the methods, results, and discussion for Paper I. The study underscores the importance of effective machine learning algorithms for sleep/wake cycles, which ideally necessitate correct data annotations over a span of several days. Although sleep diaries or EEG recordings can annotate 'time in bed', many researches exclusively rely on accelerometry. This emphasizes the imperative for valid annotation techniques. Our objective is to introduce a manual annotation method, assess its precision, and determine its consistency. Some of the details presented here were previously mentioned in the published version of Paper I[127] (see Appendix I).

Methods

Study Population

The data for this study was sourced from the SCREENS pilot trial (www.clinicaltrials.gov, NCT03788525), a two-arm parallel-group cluster-randomized trial with two intervention groups, conducted between October 2018 and March 2019[116,128]. There was no control group in this trial.

Families from the Middelfart municipality in Denmark were approached for participation if they had a child aged between 6 to 10 years living with them, out of a total of 1686 families. To qualify, the parent's screen media usage had to exceed the median of 2.7 hours per day, based on survey responses from 394 respondents. Additionally, all children in the household needed to be older than 3.9 years to ensure that sleep measurements weren't disrupted by the nocturnal awakenings typical of infants or toddlers. For a comprehensive list of inclusion and exclusion criteria, refer to Pedersen et al.[52].

The present study ultimately included data from 14 children and 19 adults. These participants weren't advised to alter their sleep or bedtime routines for the interventions. While the study focused on nightly sleep time as recorded by the EEG-based sleep staging system, any napping behavior of the participants was deemed irrelevant.

All data collection procedures were reported to the local data protection department, SDU RIO (ID: 10.391), in compliance with the Danish Data Protection Agency's regulations.

Accelerometry Recordings

Both adults and children participated in 24-hour accelerometry recordings using two triaxial accelerometers, Axivity AX3 (Axivity Ltd., Newcastle upon Tyne, UK). The Axivity

AX3 is a compact device, measuring 23 mm × 32.5 mm × 7.6 mm and weighing just 11 g. It was set with a sensitivity of ±8 g and a sampling frequency of 50 Hz. Participants wore the accelerometers at two specific anatomical locations. The first was positioned on the right hip, secured in a pocket attached to a belt around the waist, ensuring the USB connector faced outward from the body's right side. The second accelerometer was placed midway between the hip and knee on the right thigh, housed in a pocket on a belt, with the USB connector also facing away from the body. For both the baseline and follow-up, the devices were worn for a duration of one week (seven consecutive days). This duration aligns with the recommended number of days to reliably assess habitual physical activity[129].

For our study, we incorporated a total of seven distinct signal features. The criteria for classifying "lying" in the first feature are explicit: if the inclination of the hip accelerometer surpasses 65 degrees and the thigh accelerometer simultaneously identifies as "sitting" based on Skotte et al.'s activity type classification algorithm[63]. The other signal features, with the exception of "time", are directly obtained from Skotte et al.'s algorithm. These features, delineated in Table 1, concern the longitudinal axis of the body. Data derived from accelerometry undergoes processing using a window length of two seconds (60 samples) and has a 50% overlap (30 samples), ensuring a resolution of one second. The methodologies from Skotte et al. and those generating the first feature rely exclusively on the accelerometer's inclination(s). While the methodologies from Skotte et al. and the techniques generating the first feature can provide a rough estimate of time spent in bed and identify general postures, they cannot accurately pinpoint the specific moments a participant enters or leaves the bed.

EEG-based and Self-Report Sleep Recordings

Both adults and children were assessed for their sleep patterns using the ZM model DT-200 (General Sleep Corporation, Cleveland, OH, USA), Firmware version 5.1.0. This assessment was concurrent with the accelerometer recordings. At the baseline, the sleep assessment using the ZM spanned 3–4 nights, while during the follow-up, it was conducted over 3 nights.

The ZM device operates by measuring sleep through a single-channel EEG, specifically from the differential mastoid (A1–A2) EEG location, evaluated on a 30-second epoch basis. Designed for use in everyday settings, the ZM provides an objective measurement of various sleep parameters, including sleep duration, sleep stage classification, and latency to different sleep stages. The ZM's algorithms has been benchmarked against PSG in laboratory settings for both adults with and without chronic sleep issues[50,51]. Notably, the device showcased a high accuracy in distinguishing between sleep and wakefulness, with sensitivity, specificity, positive predictive value, and negative predictive values being 95.5%, 92.5%, 98%, and 84.2%, respectively. Previous findings from our lab indicate that the ZM is effectively applicable to both children and adults for multi-day measurements in real-world settings[52].

For the assessment, three electrodes (Ambu A/S, Ballerup, Denmark, type: N-00-S/25) are positioned on the mastoids (for signal) and the nape (as ground). About half an hour before their intended sleep time, participants' skin areas are cleaned with alcohol swabs,

after which the electrodes are affixed. An EEG cable connects these electrodes to the ZM device. A preliminary sensor check ensures all electrodes are correctly mounted; any issues are promptly addressed by replacing the problematic electrodes. Additionally, participants, or parents on behalf of their children, recorded their sleep and wake times daily in a dedicated sleep diary.

Annotation Software

Audacity® is a distinguished free audio editing software[130]. The genesis of Audacity can be traced back to the fall of 1999, when it emerged as an innovative project led by Dominic Mazzoni and Roger Dannenberg at Carnegie Mellon University. By May 2000, it was unveiled to the world as an open-source audio editor. Since its inception, Audacity has undergone extensive evolution. The software, developed collaboratively by the community, now boasts of hundreds of unique features, offers complete support for professional-grade 24-bit and 32-bit audio, has a comprehensive manual available in multiple languages, and has witnessed distribution in the millions. Today, a dedicated team of volunteers from various corners of the globe continues to maintain and enhance Audacity. It is disseminated under the GNU General Public License, granting everyone the freedom to utilize the software for personal, educational, or commercial endeavors.

Audacity is not limited to audio processing; it can also serve as a tool for accelerometer data analysis. This software provides researchers with the means to precisely inspect high-resolution raw accelerometer data in great detail. Users can quickly zoom in to explore deeper into specific segments of the recording, like certain patterns around bedtime, or zoom out for a broader perspective, such as data spanning a week. Furthermore, Audacity's features a sophisticated labeling function for annotating the accelerometry data. Any labels created can be saved in a separate file and subsequently incorporated into the training phase of machine learning algorithms. The depth of manual inspection for high-resolution accelerometer data that Audacity provides is, to our knowledge, matched by only a few other software options[131,132]. However, these alternatives have data import restrictions in their free versions.

Within the Audacity interface, there's the possibility of combining over 100 channels of data. This aids in the merging of distinct signal features derived from acceleration. The integration of multiple signal features is intriguing as it might enhance the visual comprehension and classification of inherent behaviors. Nevertheless, an excessive collection of signal features might obscure the precise identification of targeted behaviors.

To provide a visual perspective, Figure 2 and Figure 3 depict the Audacity interface displaying all seven signal features as cataloged in Table 1. Figure 2 offers a glimpse of a week's data, whereas Figure 3 zooms into an approximate 24-hour span, showcasing a single annotated night.

Table 1: Summary of the specific signal features utilized in Audacity for maunal annotation in-bed and out-of-bed timestamps.

Feature	Description	Values
Lying	Posture based on thigh and back	1: lying, -1: not lying
Activity	Activity type classification	1: Standing, moving, 0: Sitting, -1: Other
Time	Time categorized into four-hour windows	[-1,0.2) with step size Δ=0.2
Thigh-SDacc	Thigh longitudinal acceleration SD	-1: No movement
Thigh-Inclination	Thigh device inclination angle	Range: −180 to 180 degrees
Hip-SDacc	Hip longitudinal acceleration SD	-1: No movement
Hip-Inclination	Hip device inclination angle	Range: −180 to 180 degrees

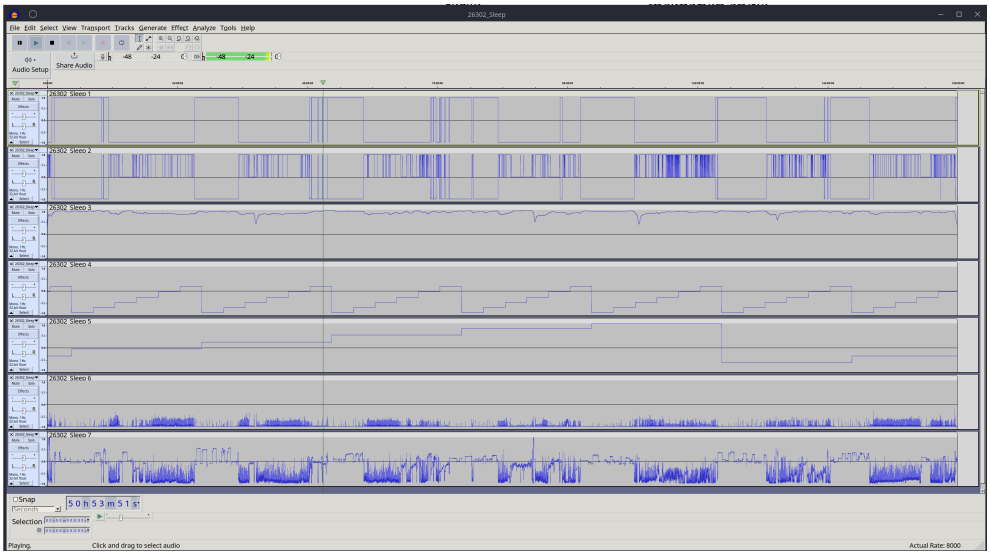

Figure 2: Screenshot of the Audacity interface showing the seven horizontal panels representing the included signal features. See Table 1 for a detailed description of the features.

Annotation Process

Three researchers, experienced in working with accelerometer data, were chosen as raters. Their proficiency ensured that they had the requisite knowledge to accurately interpret the various data channels presented to them. Each rater reviewed and labeled each wav-file, marking specific timestamps that indicated in-bed and out-of-bed activities. These annotations were then saved as individual text files. For ensuring consistency and reliability in the annotations, each wav-file underwent two rounds of labeling by each rater. Importantly, at no point during this process were the raters aware of any prior annotations, either made by themselves or their colleagues. This approach was adopted to prevent any potential biases and ensure the highest degree of objectivity in the annotations.

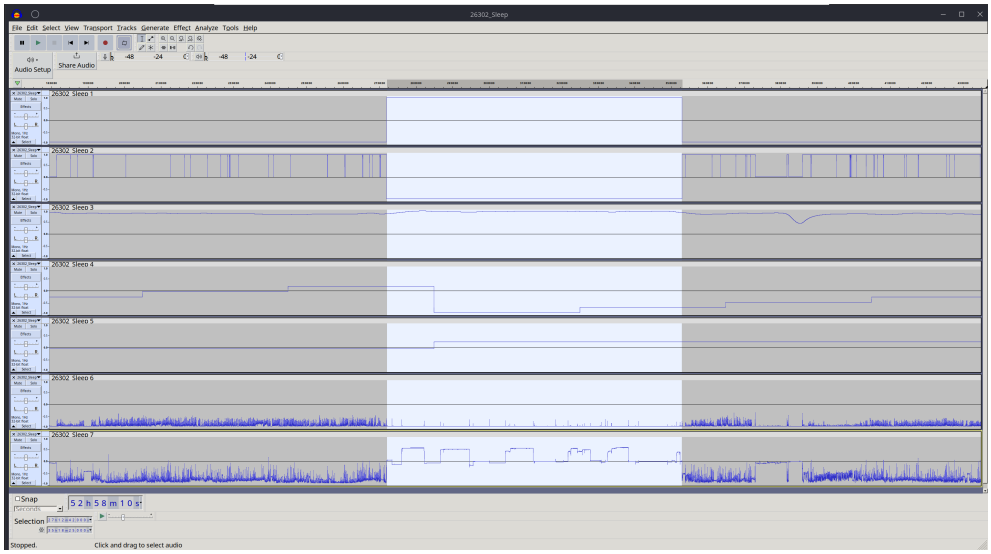

Figure 3: Screenshot of the Audacity interface when zoomed in on a single night for the labeling of the in-bed period. The seven horizontal panels represent the included signal features. See Table 1 for a detailed description of features.

Establishing the Ground Truth

The definitive ground truth for in-bed and out-of-bed time frames was obtained from the sleep staging data derived from the ZM device. This was established by identifying the first and last events at night that did not present any sensor-related issues. Nights where the ZM detected sensor problems, either at the onset or conclusion of the recording, were excluded from further consideration. Such sensor issues typically arise due to high impedance because of inadequate attachment of electrodes. To maintain accuracy in data collection, all participants were meticulously instructed to affix the ZM before bedtime and then activate it precisely at their bedtime and to detach it upon waking. These timestamps were then utilized as the ground truth for the study.

Statistics

For continuous variables, the descriptive characteristics were computed using medians and interquartile ranges. Meanwhile, categorical variables were assessed based on their proportions. To offer a clear distinction, the characteristics for children and adults are presented separately.

In assessing the consistency of our findings, we utilized the intraclass correlation coefficient (ICC) alongside the Bland–Altman analysis. All timestamps were transformed into seconds past midnight to provide a continuous measure for the ICC analyses. Recognizing that the human raters were sampled from a broader population, we used a two-way random-effects model when assessing inter-rater reliability between different human raters. Here, ICC(2,k) was chosen, reflecting the absolute agreement between

multiple rater's ratings[133]. However, when comparing human raters against the ZM or sleep diaries, a two-way mixed-effects model was used. In this context, the human raters were treated as random effects, while the ZM and sleep diaries were treated as fixed effects. The corresponding ICC that represents this model is known as ICC(3,k)[133]. For intra-rater agreement, a two-way mixed effects models was employed treating human raters as random effects and occasion (test/retest) as fixed effects. We adopted the ICC(3,k) to estimate the agreement of each individual rater's ratings across occasions[134]. In general, the ICC serves as a more nuanced tool than simple correlation; it goes further by evaluating the alignment in magnitude between two datasets, serving as a robust metric for consistency between methodologies. The interpretation of ICC values were as follows:

$$\text{ICC} < 0.5 \text{ indicates poor agreement}$$

$$0.5 \leq \text{ICC} < 0.75 \text{ indicates moderate agreement}$$

$$0.75 \leq \text{ICC} < 0.9 \text{ indicates good agreement}$$

$$\text{ICC} \geq 0.90 \text{ indicates excellent agreement}$$

In this paper, the ICC values are presented as ICC (95% CI) and is interpreted based on their 95% confidence intervals, for example, a CI of 0.83-0.94 indicates "good" to "excellent" agreement, while a CI of 0.92-0.99 is solely "excellent" as even the lowest value surpasses 0.9. By doing so, we adhere to recommended guidelines as presented by Koo et al.[135]. The ICCs were calculated using the R package psych[136]. The Bland–Altman analysis evaluates the agreement between two measurement techniques[137]. It calculates the mean difference between the two methods (representing bias) and establishes the limits of agreement. A positive mean difference indicates an underestimation, meaning the in-bed or out-of-bed timestamp is earlier in the day relative to the ZM. Conversely, a negative difference denotes a later timestamp compared to ZM. To visually present this agreement, we used probability density distribution plots, illustrating the symmetry between the methods. All statistical analyses were conducted using R (version 4.0.2) and RStudio (version 1.1.456).

Results

Descriptive characteristics of the included subjects of the current study are reported in Table 2.

Table 2: Descriptive characteristics of the study participants. ISCE: International Standard Classification of Education

Characteristic	
Children	
N	14
Gender (% female)	28.6
Age (years)	9 (7–10)
Adults	
N	19
Gender (% female)	57.9
Age (years)	42 (39–46)
ISCE	
0–3 (%)	36.8
4–6 (%)	47.4
7–8 (%)	15.8

Intraclass Correlation Coefficient Analyses

The ICCs analysis, as displayed in Table 3, showed excellent agreement between the ZM and the averaged manual annotations made by the three human raters across all comparisons. ICC values of the baseline comparisons (covering 94 nights) were consistently 0.98 with the lower limit of the 95% confidence intercal only dipped as low as 0.96 indicating excellent agreement. Similarly, all follow-up comparisons (encompassing 54 nights) showed ICCs above 0.95 with the lowest 95% confidence interval scoring 0.92 for the second round of "to-bed" annotations ensuring excellent agreement.

Table 3: Intraclass correlation coefficients between the ZM and the three human raters. Values are ICC (95% CI).

	Baseline (N = 94)		Followup (N = 54)	
	Round 1	Round 2	Round 1	Round 2
To bed	0.98 (0.98; 0.99)	0.98 (0.96; 0.98)	0.96 (0.94; 0.98)	0.95 (0.92; 0.97)
Out of bed	0.98 (0.97; 0.99)	0.98 (0.96; 0.98)	0.98 (0.97; 0.99)	0.97 (0.95; 0.98)

The weakest agreement between self-report- and ZM-determined in-bed periods were observed for the "to-bed" timestamp on the follow-up data yielding a lower limit of the 95% confidence interval of 0.94 still indicating excellent absolute agreement (see Table 4).

Table 4: Intraclass correlation coefficients between self-report and the ZM. Values are ICC (95% CI).

	Baseline (N = 94)	Followup (N = 54)
To bed	0.98 (0.98; 0.99)	0.96 (0.94; 0.98)
Out of bed	0.98 (0.97; 0.99)	0.98 (0.96; 0.99)

Assessing the agreement between the three human raters when annotating timestamps for 'to bed' and 'out of bed' events, the ICC values reflected good to excellent agreement between the raters across both rounds and timestamps. Specifically, the lower bounds of the 95% confidence intervals dipped below 0.9 (ICC > 0.9 indicating excellent agreement) for round 1 and round 2 "to-bed" on the baseline data, with the least value being 0.88 (see Table 5).

Table 5: Intraclass correlation coefficients between the three human raters. Values are ICC (95% CI).

	Baseline (N = 110)		Followup (N = 62)	
	To Bed	Out of Bed	To Bed	Out of Bed
Round 1	0.91 (0.88; 0.94)	0.93 (0.9; 0.95)	0.94 (0.9; 0.96)	0.97 (0.96; 0.98)
Round 2	0.92 (0.89; 0.94)	0.97 (0.96; 0.98)	0.97 (0.95; 0.98)	0.98 (0.98; 0.99)

Across baseline and follow-up and on both to-bed and out-of-bed timestamps, each rater displayed good to excellent test-retest agreement, with the lower limits of the 95% confidence interval of hte ICCs values ranging from 0.86 to 0.99. Notably, while Raters 1 and 3 demonstrated a minor dip in their baseline to-bed agreement compared to out-of-bed measures, Rater 2 showed lower agreement on the follow-up to-bed timestamp compared to out-of-bed (refer to Table 6).

Table 6: Test–retest intraclass correlation coefficients between the first and second round of manual annotations. Values are ICC (95% CI).

	Baseline (N = 110)		Followup (N = 62)	
	To Bed	Out of Bed	To Bed	Out of Bed
Rater 1	0.91 (0.87; 0.94)	0.98 (0.98; 0.99)	0.96 (0.94; 0.98)	1 (0.99; 1)
Rater 2	0.97 (0.96; 0.98)	0.91 (0.87; 0.94)	0.91 (0.86; 0.95)	0.99 (0.98; 0.99)
Rater 3	0.91 (0.87; 0.94)	0.96 (0.94; 0.97)	0.98 (0.97; 0.99)	0.98 (0.97; 0.99)

Bland-Altman Analyses

Table 7 outlines the Bland–Altman analyses comparing both human raters and self-report against the ZM in-bed and out-of-bed timestamps. The bias observed for human raters against the ZM fluctuates between -6 minutes to 5 minutes, suggesting a relatively narrow range of mean differences across evaluations. Comparatively, the self-report's bias against ZM is slightly more constrained. The ranges of the limits of agreement remained fairly consistent irrespective of the method being contrasted with ZM at -43.81 to -21.3 minutes for the lower LOAs and 20.87 to 35.85 minutes for the upper LOAs. This

uniformity in limits of agreement suggest similar agreements of both manual annotations and self-reports when compared with the ZM in-bed periods.

Table 7: Bland–Altman analysis comparing manual annotation and self-report to ZM measurements, with all data presented in minutes.

Method	Bias (95% CI)	Lower LOA (95% CI)	Upper LOA (95% CI)
Baseline, to bed, n = 94			
Manual, round 1	3.02 (-0.44; 6.47)	36.07 (30.15; 42)	-30.04 (-35.96;-24.12)
Manual, round 2	0.48 (-2.42; 3.39)	28.27 (23.29; 33.24)	-27.3 (-32.28;-22.32)
Self-report	1.23 (-1.57; 4.03)	28.02 (23.21; 32.82)	-25.56 (-30.37;-20.76)
Baseline, out of bed, n = 94			
Manual, round 1	0.53 (-2.34; 3.4)	27.96 (23.05; 32.88)	-26.9 (-31.82;-21.99)
Manual, round 2	0.98 (-1.47; 3.43)	24.45 (20.24; 28.66)	-22.49 (-26.7;-18.28)
Self-report	-2.79 (-5.26; -0.32)	20.87 (16.63; 25.11)	-26.45 (-30.69;-22.21)
Follow-up, to bed, n = 54			
Manual, round 1	-6.08 (-11.34; -0.83)	31.64 (22.61; 40.67)	-43.81 (-52.84;-34.77)
Manual, round 2	-0.4 (-5.3; 4.51)	34.8 (26.37; 43.23)	-35.6 (-44.03;-27.17)
Self-report	0.77 (-4.08; 5.62)	35.59 (27.25; 43.93)	-34.06 (-42.4;-25.72)
Follow-up, out of bed, n = 54			
Manual, round 1	4.95 (0.65; 9.25)	35.85 (28.45; 43.25)	-25.95 (-33.35;-18.55)
Manual, round 2	2.57 (-0.76; 5.89)	26.44 (20.72; 32.15)	-21.3 (-27.02;-15.59)
Self-report	0.56 (-3.62; 4.74)	30.57 (23.39; 37.76)	-29.45 (-36.64;-22.26)

Density Plots

Figure 4 presents the probability density distribution of the differences between the "to-bed" and "out-of-bed" timestamps, comparing manual annotations and self-reports to the ZM. These plots offer a visual illustration of the bias and spread around zero, showcasing how manual annotations and self-reports diverge from ZM, as previously highlighted[104].

Discussion

In this study, we introduced a methodology for manually annotating periods spent in bed using accelerometry data. The accuracy of this method was evaluated across multiple raters, within raters (test/retest) and then compared to in-bed and out-of-bed timestamps as detected with the ZM. Furthermore, the comparison of self-reported in-bed and out-of-bed timestamps as obtained from prospective sleep diaries was tested against the ZM. All comparisons were made using ICC and Bland-Altman analyses. When examining the limits of the 95% confidence interval of the ICC analyses, we found several noteworthy results. First, the method exhibited good-to-excellent inter-rater agreement between human raters. Second, intra-rater agreement of the human raters also showed good to excellent agreement across all three raters between their first and second rounds of annotations. Third, when comparing human raters to the ZM, the ICC indicated excellent

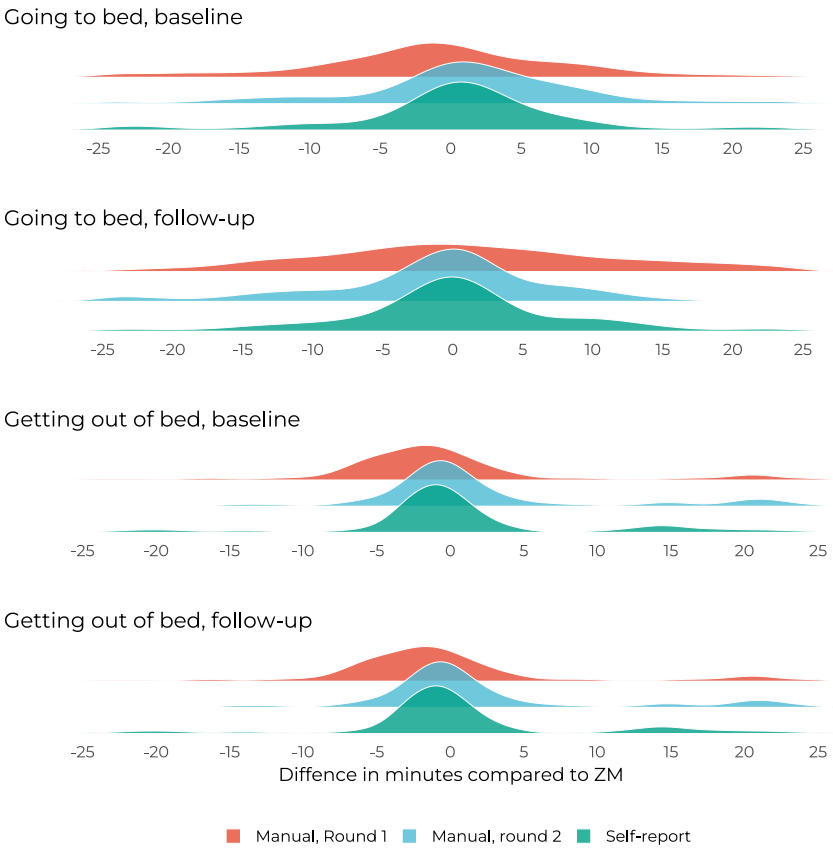

Figure 4: Density distributions of timestamp differences: Manual annotations vs ZM and Self-report vs ZM for in-bed and out-of-bed times.

agreement. Fourth, comparing in-bed and out-of-bed timestamps of self-reported vs ZM, we also observed excellent agreement between the two. Additionally, Bland-Altman analysis indicated that the mean bias between both the manual annotations and self-reported sleep times compared to ZM was within a range of ±6 minutes, with LOAs not exceeding ±45 minutes. Probability density distribution plots further substantiated these findings, showing comparable symmetry, spread around zero, and positioning of outliers when the manual annotations and self-reported sleep times were compared to ZM.

The high accuracy observed between the ZM and the prospective sleep diaries in this study can be attributed to the sleep diaries being synchronized with ZM. Having participants manually start and end the ZM recordings every morning and evening enhances their ability to accurately recall their times of going to bed and getting out of bed. This minimizes the usual discrepancies often seen between objectively and subjectively measured sleep durations[138]. If participants had been instructed to log their sleep using a retrospective sleep questionaire or without these protocol anchors of the ZM, we would expect to see less strong agreement between the manual annotations and sleep diaries compared to the ZM timestamps.

Compared to ZM, our study found that the manual annotation of in-bed and out-of-bed timestamps was more prone to errors when estimating the time of going to bed. This was expected as the issue mainly arose from raters having difficulties in differentiating between inactive behaviors before bed time and actual bed time. Despite this limitation, the manual annotations displayed reassuring accuracy, especially considering the limited formal instructions provided to the raters. This ease of use was aided by the specific signal features we selected for study in Audacity, as evidenced by the excellent ICC agreement scores between the raters. Interestingly, our findings suggest a learning curve for the raters, as evidenced by the narrower LOAs and the density plots in the second round of manual scoring. These indicators suggest that additional rounds of scoring could further improve result consistency or that preliminary training could be beneficial for the raters.

While other tools are available for annotating time series data, including Label Studio[132], Visplore[131], our research determined that Audacity was sufficiently suited for our specific requirements. Label Studio is open-source and free, making it a viable alternative; however, when dealing with extensive datasets, such as week-long accelerometer data comprising over 100 million entries, it might face challenges due to browser restrictions, and backend configuration. On the other hand, while Visplore is tailored for visual exploration of time series data, its free version comes with a data import limitation of 50 MB and offers only a subset of its features.

We deliberately tailored our feature selection to prevent overwhelming the raters with superfluous information, choosing a concise yet effective set of features rooted in domain expertise. This strategy could be adapted for annotating other activities, such as walking, which would, however, necessitate a different feature set. The human raters in this study gained valuable insights despite the absence of explicit guidelines for data annotation, highlighting the intuitive nature of the method. It's important to note that labeling data inherently involves a certain level of understanding of human behavior in accelerometry data. If such labels could be accurately determined based on a set of formal rules (i.e., a heuristic algorithm), it raises the question of whether training a machine learning

model would even be necessary. Examining the impact of varying feature sets could yield further insights that would streamline the manual annotation of accelerometer time series data.

To date, many studies comparing actigraphy and self-report methods to PSG or EEG-based methodologies have primarily focused on evaluating aggregate sleep measures like total sleep time, wake after sleep onset, sleep latency, and sleep efficiency[139,140]. These measures inherently incorporate aspects like sleep onset and wake onset times, akin to the "to-bed" and "out-of-bed" timestamps of our study. However, the precision of these specific time points has been scarcely assessed, making direct comparisons with our study difficult. One study by van Hees and colleagues reported mean absolute errors of 39.9 minutes for sleep onset time and 29.9 minutes for wake-up time, with 95% limits of agreement surpassing ±3 hours when comparing the HDCZA algorithm to PSG[104]. It's worth noting that determining exact timestamps of specific events is inherently more challenging than summarizing broader measures like total sleep time. The former demands high precision, while the latter averages variations across longer periods, naturally minimizing potential discrepancies. Our research underscored the strengths of manual annotations. One consistent observation was their reliability across a broad age and gender spectrum, which was reflected in our diverse sample that included both children and adults from both genders. This highlights the adaptability of manual annotations, given that sleep behaviors are often influenced by developmental stages and gender-specific factors. Furthermore, despite being more labor-intensive, manual annotations seem to offer superior precision, particularly when identifying exact moments compared to the HDCZA algorithm which seem to struggle with nuances that human raters more easily detect. Therefore, it seems that manual annotating the in-bed and out-of-bed timestamps is better at delivering consistent results across varied groups which indicates its potential for broader applicability. This consistency is crucial when extending findings, especially in studies focusing on typical sleep patterns.

Identifying sleep periods as opposed to merely lying down in bed is a critical aspect of 24-hour behavior profiling. Traditional studies on sleep detection often rely on participants to self-report their time in bed, sleep onset, and wake-up times[42,75,126]. However, the manual annotation methodology offers an alternative that not only reduces the burden on participants but also mitigates the recall bias inherent in self-reported measures. This method of manual annotations can be easily applied to free-living data, making it incredibly versatile for various applications beyond sleep detection. For instance, the manual annotations is useful for annotating non-wear time, manually synchronizing clocks across different devices, and examining the validaty of raw data and more. Its applicability also extends to multi-channel data, providing a comprehensive overview that can incorporate variables like orientation from gyroscopic data, temperature, battery voltage, and light. Audacity stands out for its capability to handle large multi-channel data effortlessly. Researchers can quickly zoom to any resolution and scroll through time without experiencing lag, which makes it an ideal tool for adding labels. This fluidity in workflow suggests that Audacity could become a standard tool for researchers working with raw data for labelling purporses and in relation to machine learning applications.

For years, the transition from raw sensor data to operational predictive models has relied on labeled data. Despite this, no previous research has offered a insights into the precision of manual annotations compared to self-report measures and in-bed periods

as determined by an EEG-device which allows researchers to optimally utilize their available accelerometry data. Our study demonstrates that with a careful selection of features, manual annotation for identifying in-bed and out-of-bed timestamps can yield results comparable to those achieved with other methodologies. However, it's crucial to clarify that we are not advocating that manual annotations should replace more established techniques for sleep estimation, such as EEG or tracheal-sound-based methods, in ongoing studies. Instead, manual annotations can be valuable as a post-hoc procedure to enrich existing datasets with an additional measure of human behaviors.

This study boasts several strengths, notably the continuous, multi-day data collection of accelerometry, sleep diary, and ZM recordings carried out in participants' homes, which provides high-quality free-living data. However, the study is not without limitations. One such limitation pertains to rater generalizability. The three manual raters were all experienced working with accelerometry data, and as such, their proficiency might not be representative of the broader population of potential raters, possibly affecting the replicability of our findings with less experienced individuals. Nevertheless, given the minimal pre-briefing instructions for labeling raw data, we believe this methodology should be generalizable to other researchers working with accelerometer data. Another concern is the challenge of recording true free-living behavior using participant-mounted devices like ZM, as wearing such a device during sleep could affect participants' natural behavior, thereby posing a study limitation. Finally, the study did not consider napping behavior; its focus was solely on in-bed periods as it relates to circadian rhythms. As such, future research is required to validate the utility of this manual annotation methodology for detecting naps.

Paper II: Generalizability and Performance of Methods to Detect Non-Wear With Free-Living Accelerometer Recordings

This segment of the thesis encompasses the methods, results, and discussion for Paper II. Despite advancements in sensor technology and software development, the accurate classification of non-wear time in raw accelerometer data remains a challenge. This raises an important question: which heuristic algorithm or machine-learning model can best classify non-wear time when analyzing unseen accelerometer data? To address this, three datasets were generated, each comprising raw accelerometer data manually annotated for wear and non-wear times and inclusive of surface skin temperature measurements. This study specifically aimed to train three decision tree models using data from thigh and hip-worn accelerometers to classify non-wear time. Furthermore, the importance of surface skin temperature was evaluated, and the potential advantages of limiting the number of features in the decision tree model were explored. Lastly, the comparative performance of the developed decision tree models against basic heuristic algorithms and recently developed random forest and convolutional neural network models was assessed. Some of the details presented here were previously mentioned in the published version of Paper II[141] (see Appendix II).

Methods

Reference Methods Overview

A total of four additional non-wear classification methods were incorporated to assess generalizability and contrast their performance with the three developed decision tree models. These pre-existing methods were deliberately chosen to encompass a range of methodological flexibility. This selection ensured representation from the simplest and most commonly used techniques to the latest and most complex ones.

1. Consecutive Zeros-Algorithm (`cz_60`): Over the years, there have been various consecutive zero-algorithms designed for accelerometer data, with the aim of identifying non-wear periods within predefined timeframes, such as 30-, 60-, or 90-minute intervals[105,110,142]. In research by van Hees et al.[143], the potential of a simple summary measure derived from raw triaxial accelerometer data to aid in the estimation of PA-related energy expenditure in both pregnant and non-pregnant women was explored. The study involved 108 women from Sweden and 99 women from the United Kingdom who wore a triaxial GENEA accelerometer for durations of 10 and 7 days, respectively. The researchers developed an algorithm to

discern wear and non-wear time, basing their estimates on the standard deviation and value range of each accelerometer axis over 30-minute intervals. Intervals were designated as non-wear time if the standard deviation was below 3.0 mg (1 mg = 0.00981 ms^{-2}) for at least two of the three axes, or if the value range was under 50 mg for at least two of the three axes. In a subsequent study by van Hees et al.[144], the interval length was extended to 60 minutes to reduce the likelihood of misidentifying sedentary periods as non-wear time. Furthermore, a 15-minute sliding window was introduced to account for overlapping episodes and to pinpoint non-wear episode boundaries more accurately. Another method utilizes a 135-minute interval with adjusted hyperparameters, as introduced by Syed et al.[145]. In our study, we adopted a straightforward approach to this concept. Using Actigraphy counts, we identified periods of no movement that registered zero counts for at least 60 continuous minutes. Notably, these Actigraphy counts operate with a deadband set at 68 mg, which denotes the minimum detectable acceleration threshold.

2. Heuristic Algorithm (`heu_alg`): As detailed by Rasmussen and colleagues[116], this algorithm merges raw acceleration data with surface skin temperature measurements. Non-wear time is determined for periods surpassing 120 minutes with accelerations less than 20 mg. For durations between 45 to 120 minutes, non-wear is identified if the temperature falls below a personalized non-moving temperature threshold. Additionally, the algorithm can spot non-wear periods ranging from 10 to 45 minutes, but only if these intervals end within the anticipated awake hours (06:00 AM to 10:00 PM).

3. Random Forests Model (`sunda_RF`): Sundararajan et al.[100] delineated a non-wear classification technique grounded in a random forest ensemble model. This model was informed by raw accelerometer data derived from 134 participants aged between 20 to 70 years. These subjects were fitted with an accelerometer on their wrist for a singular overnight PSG recording session. The ground truth labels for non-wear periods were anchored in the assumption that the accelerometer was always worn during the PSG recording. Any epoch with a standard deviation in the acceleration signal exceeding 13.0 mg outside the PSG recording time period was classified as wear time. The model utilized 36 predictors, and a nested cross-validation method was employed both to ascertain the model's generalization capability and to tune its hyperparameters.

4. Deep Convolutional Neural Network (`syed_CNN`): This method, introduced by Syed et al.[119], employs a unique approach. It's built upon a deep convolutional neural network that diverges from traditional techniques. Initially, all potential non-wear episodes are discerned using a standard deviation threshold. However, instead of examining the acceleration within these intervals, the focus shifts to the signal shape of the raw acceleration immediately before and after a non-wear episode. Through the convolution neural network, the method discerns non-wear periods by detecting the moments when the accelerometer is removed and reattached. For our study's purposes, we chose a window length of 10 seconds on each side of the identified non-wear episode, as this yielded the most accurate results. The training dataset that informed the CNN consisted of data from hip-mounted accelerometers

worn by 583 participants. These individuals ranged in age from 40 to 84 years, with an average age of 63 years and a standard deviation of 10.

Data Sources

Data for the current study were sourced from the PHASAR study[146] and an in-house validation study with both studies utilizing the Axivity AX3 accelerometer (Axivity Ltd., Newcastle upon Tyne, UK) to record raw acceleration data along with surface skin temperature. The device, weighing a mere 11 g and with dimensions of 23 mm × 32.5 mm × 7.6 mm, measures acceleration in gravity units (g) across three axes (vertical, mediolateral, and anteroposterior). The sampling frequencies were collected at 50 Hz for the PHASAR study and 25 Hz for the in-house study. However, all recorded data from both studies were uniformly resampled to 30 Hz.

The PHASAR study involved a representative sample of over 2000 school-aged children from 31 public schools in Denmark. The study, conducted between 2017 and 2018, captured data from 1,315 boys (49%) and 1,358 girls (51%), aged between 8.1 to 17.9 years (mean age = 12.1, SD = 2.4). Accelerometers were placed at two specific anatomical sites: the right hip and midway on the right thigh. They were worn for a recommended seven consecutive days to reliably estimate habitual physical activity. For this analysis, data from 64 randomly selected participants from the PHASAR cohort were used. A dataset indicating ground truth non-wear time was created via manual annotation, a method elaborated in Paper I. Essentially, non-wear periods were determined by visually examining raw accelerations coupled with skin temperature readings. True non-wear episodes with specific start and end times were manually labelled in each dataset and were utilized as reference labels in subsequent analyses.

The in-house validation study consisted of accelerometer data from 42 youth athletes, evenly split between boys and girls, aged 14.5 to 16.4 years (mean age = 15.4, SD = 0.4 years). These athletes, part of a specialized talent program in the Region of Southern Denmark, wore the Axivity accelerometer on their non-dominant wrist for 14 consecutive days. This study was initiated in the spring of 2021. A dataset with ground truth non-wear time was created mirroring the dataset drawn from the PHASAR study, including all 42 participants.

The PHASAR study was reviewed by the Regional Committee on Health Research Ethics for Southern Denmark (ID: S-20170031) and was determined not to require an ethics review, as per Danish regulations, which mandate only biomedical research or risk-involved studies to undergo a formal ethics review. Documentation regarding this decision is available upon request from the corresponding author. Conversely, the in-house validation study received an ethical approval waiver from the Research & Innovation Organization and the legal department of the University of Southern Denmark. All participants, or their legal guardians, provided written informed consent for both studies, which adhered to the Danish Data Protection Agency (2015-57-0008) standards and globally recognized guidelines like the Declaration of Helsinki.

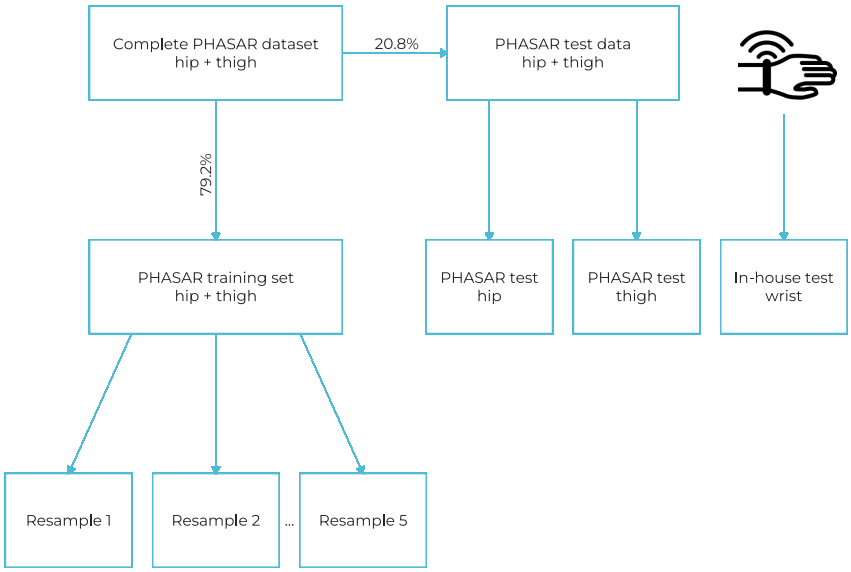

Figure 5: Flowchart illustrating the division of the PHASAR dataset into training and testing datasets. On the left, boxes represent 79.2% of the PHASAR data allocated for training across five-fold resamples. In the middle, boxes represent 20.8% of the PHASAR data delineated for testing, specifically marking the hip and thigh data. The box on the right-hand side signifies our in-house test dataset obtained from wrist-worn devices.

Development of Decision Tree Models

For our decision tree models, we sourced 12 features from the raw PHASAR accelerometer data, which encompassed elements like temperature, time of day, indicators for device placement, day of the week, and moving average statistics (detailed in Table 8). These moving average metrics were collated in 10-second increments. To train the model, we utilized 79.2% of the PHASAR data, incorporating data from both hip- and thigh-worn devices (as shown in Figure 5). We made certain that data from individual participants was exclusively allocated to either the training or test datasets. This strategy was to ensure that the model could effectively generalize to unfamiliar data, rather than overfitting to specific participant data. During the tuning phase, to boost model accuracy and avoid overfitting, we opted for a five-fold cross-validation approach. This process entailed refining several hyperparameters, such as the tree's depth, its cost-complexity, and the minimum amount of data points necessary in a node for it to split further. To effectively explore the hyperparameter space, we employed Latin hypercube sampling. This method systematically divides the parameter range into segments, randomly drawing a value from each segment, resulting in a well-distributed set of parameter combinations. In our case, we established a 10-level parameter grid to search the hyperparameter space. The F1 score was treated as the optimization metric.

Following this procedure, we introduced three distinct model variations:

1. A full-scope model (`tree_full`) incorporating every feature.

2. A refined model (`tree_imp6`) centered on the six most crucial feature, as established by permutation feature importance (Figure 6).
3. A model excluding surface skin temperature as a feature (`tree_no_temp`).

In sum, our methodology generated 50 distinct models for each decision tree variant. Given our data's distribution - 55.8% wear time compared to 44.2% non-wear time - there was no need to adopt synthetic minority oversampling methods like SMOTE or other balancing techniques.

Table 8: Features extracted from the raw sensor signals.

Predictor	Description
Weekday	Day of week ([1:7])
time_day	Time of day (milliseconds)
macc_x	Moving average of the z axis acceleration
macc_y	Moving average of the y axis acceleration
macc_z	Moving average of the z axis acceleration
sdacc_x	Moving average of the standard deviation on the x axis acceleration
sdacc_y	Moving average of the standard deviation of the y axis acceleration
sdacc_z	Moving average of the standard deviation of the z axis acceleration
Sdmax	Maximum standard deviation
Incl	Inclination angle of the device in relation to the direction of the gravitational force
Temp	Surface skin temperature (degrees Celsius)

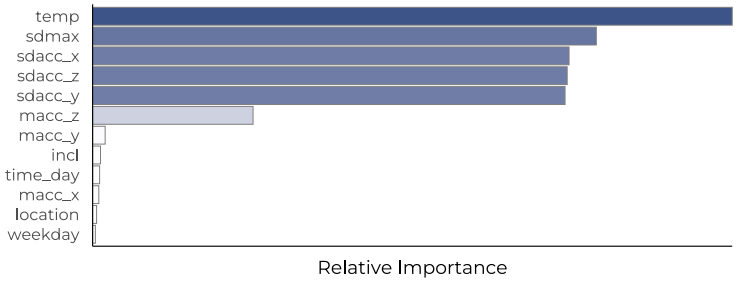

Figure 6: Permutation importance plot depicting the relative importance of predictors in the full decision tree model (`tree_full`). The top six predictors informed the `tree_imp6` model, while a third model, `tree_no_temp`, was trained using all predictors except temperature.

Statistics

Classification performance was evaluated against a ground truth test dataset, encompassing over 7 million epochs of 10 seconds each, derived from 104 unique subjects. Accurate identification of true non-wear time and true wear time yielded true positives (TP) and true negatives (TN), respectively. These correct classifications are essential for ensuring the algorithm's precision in determining non-wear time. Conversely, misclassifications, where true non-wear time is identified as wear time or vice versa, contributed to false negatives (FN) and false positives (FP). By analyzing these 10-second acceleration data

intervals against the ground truth, a confusion matrix was constructed. From this matrix, the following performance metrics were extracted:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$sensitivity = \frac{TP}{TP+FN}$$

$$precision = \frac{TP}{TP+FP}$$

$$F1\ score = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}$$

The F1-score, which represents the harmonic mean of precision and sensitivity, serves as a robust metric for evaluating classification performance, with higher F1-scores indicating better classification efficacy. Additionally, we utilized permutation feature importance to assess the contribution of each feature in the full decision tree model. This method assesses the relative importance of each feature by evaluating the decrease in model accuracy when the data for that particular feature is randomized. A significant drop in performance upon randomizing a feature suggests an important role for that feature in the model.

The decision tree models were trained using data from devices worn on the hip and thigh. On the other hand, the random forest model by Sundarajan et al. was developed using data from wrist-worn accelerometers, while the convolutional neural network by Syed et al. was trained on data from hip-worn devices. To assess the generalizability of all machine learning models, each underwent external validation. This process entailed testing them on datasets representing populations and wear locations different from those on which they were initially trained, enabling an assessment of their generalizability across varied anatomical positions and age spans (see Figure 5).

For all analyses and model development, we utilized R (version 4.1.2, Bird Hippie) and RStudio (version 2021.9.1.372, Ghost Orchid). The machine learning tasks were primarily facilitated by the Tidymodels[147] suite of packages, and we used the rpart[148] package as the engine for our decision tree algorithms.

Results

In our datasets spanning three wear locations, there were 1,598 non-wear time episodes. Of these, 1,148 episodes (or 71.8%) lasted 60 minutes or more, with an average duration of about 13 hours (794 minutes with a standard deviation of 1,142 minutes). In contrast, episodes lasting 60 minutes or less made up 28.2% (450 episodes) with an average duration of 26.4 minutes (SD = 16.4). Interestingly, the briefest episodes (less than 60 minutes) made up just 1.3% of the total non-wear time across all wear locations (refer to Table 9). Figure 7 depicts the frequency distribution for episodes shorter than 60 minutes and those 60 minutes or longer. The PHASAR dataset (hip and thigh) showed a bimodal distribution for shorter episodes, with longer episodes peaking around 10 hours. For the

in-house wrist-worn dataset, shorter episodes displayed a more uniform distribution with a peak around 15 minutes in duration, while longer episodes were significantly right-skewed.

Table 9: Overview of non-wear episodes grouped in short and long non-wear episodes, min = minutes, hrs = hours.

Wear location	Mean (min)	Cumulated (hrs)	Proportion (%)
<60 minutes			
hip	28	53	1.1
thigh	25	66	1.4
wrist	27	78	1.3
≥60 minutes			
hip	828	4663	98.9
thigh	776	4672	98.6
wrist	782	5853	98.7

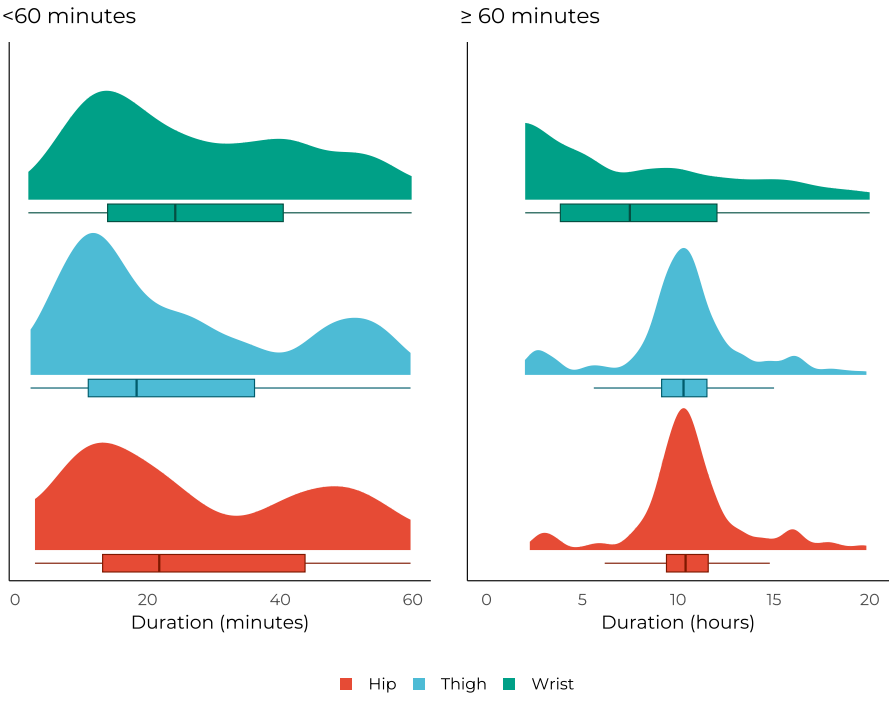

Figure 7: Distribution of the length of the non-wear episodes across hip, thigh, and wrist data. Distributions are shown for episodes shorter than 60 min and longer than 60 min.

Classification Performance

In assessing classification performance, Figure 8 visually contrasts the results from machine-learned models and rule-based algorithms against the ground truth non-wear time, which is highlighted with a light grey background. This visualization underscores that while tree-based models tend to be precise, they can also be unpredictable. On the other hand, threshold-based methods, such as `heu_alg` and `cz_60`, offer more consistency. Notably, both `cz_60` and `heu_alg` algorithms fall short in identifying shorter non-wear episodes.

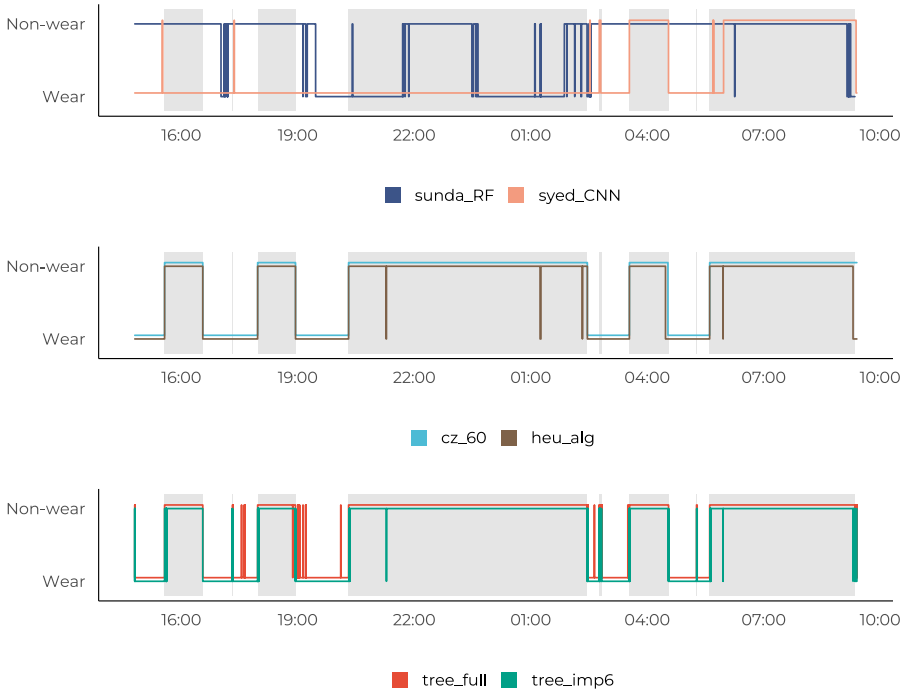

Figure 8: Visual example of a single day of the output of non-wear detection models and algorithms for a random person from the in-house wrist dataset. The grey shade is ground-truth non-wear time. `syed_CNN`, `cz_60`, and `tree_full` are vertically offset for easier interpretation.

Figure 9 compiles performance metrics from all methods evaluated in this study on all non-wear periods. The `tree_full`, `tree_imp6`, cv_60, and `heu_alg` all achieved similar performance, achieving an accuracy and F1 scores above 90% on both the thigh, hip, and wrist data. These methods showcased a uniform performance across wear location with exception of the three_full model eliciting a slight drop in performance on the wrist data. On the other hand, some models demonstrated varied performances when transitioning from one dataset to another. For instance, the `tree_no_temp` and the `syed_CNN` model's F1 score on the hip and thigh datasets hovered around 65% and 68%, respectively, but when applied to the wrist dataset, there was a noticeable dip to 47% and 59%, respectively.

The **sunda_RF** model, however, showed an upward trajectory, moving from an accuracy of approximately 57% on the hip and thigh datasets to a significantly improved 93% on the wrist dataset.

Further, Figure 10 zeroes in on performance metrics for episodes 60 minutes or shorter. The consecutive zeros algorithm was unable to detect any non-wear, a result absent from the figure. Syed et al.'s deep learning model underperformed, detecting a mere 1–2% of all non-wear time, leading to F1 scores below 5%. Although the **heu_alg** algorithm boasted high precision, its lackluster sensitivity resulted in F1 scores spanning from 12% to 16% across wear locations. The random forest model displayed average results for thigh and wrist data but faltered with hip data, recording F1 scores of 46%, 57%, and 8%, respectively. Among the trio of decision tree models, the one leveraging the six most important features outshone the rest, with F1 scores between 72% and 79%. Meanwhile, the decision tree model encompassing all predictors faced challenges with hip data due to a 23% sensitivity score. The decision tree model excluding the surface skin temperature exhibited commendable precision; however, its low sensitivity culminated in F1 scores ranging from 45% to 57%.

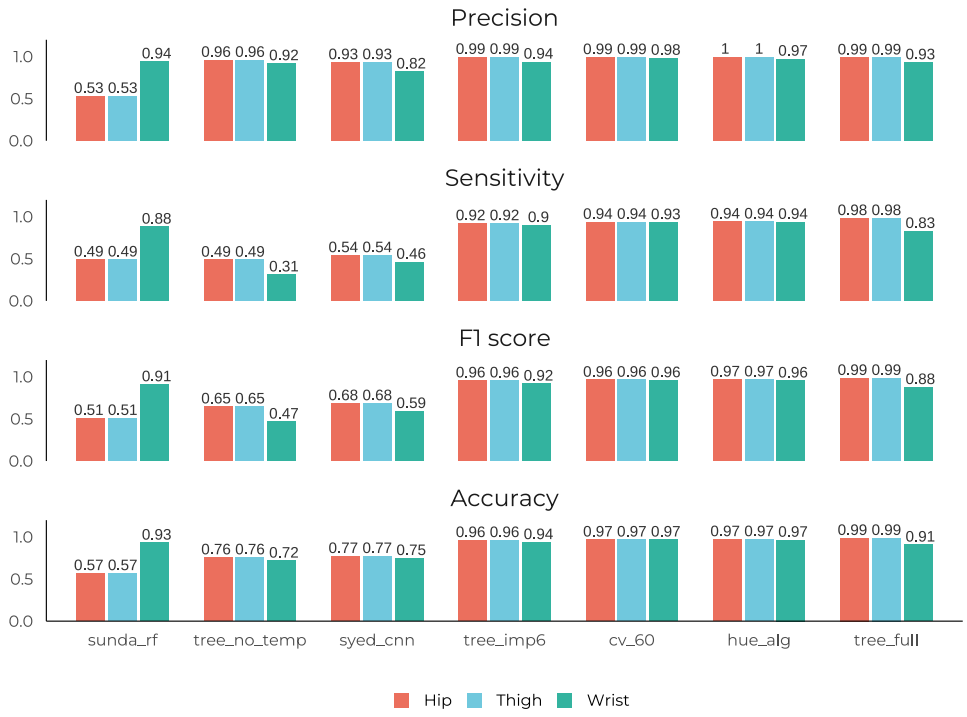

Figure 9: Classification performance metrics for all non-wear episodes as assessed by all included methods for classifying non-wear time. Metrics are displayed for three different ground-truth datasets: hip-worn, thigh-worn, and wrist-worn raw accelerometer data.

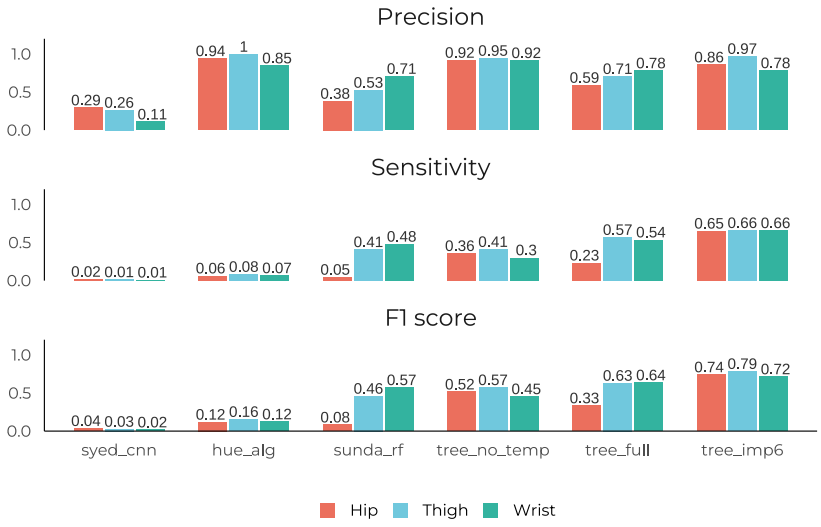

Figure 10: Classification performance for episodes no longer than 60 min in length. Metrics are shown for the three different gold-standard datasets: hip-worn, thigh-worn, and wrist-worn raw accelerometer data.

Discussion

In this study, we evaluated various methods for classifying non-wear episodes in free-living accelerometer data, focusing on all non-wear episodes and those shorter than 60 minutes. Our findings showed that the simplest methods, specifically `cz_60` and `heu_alg`, excelled in identifying non-wear episodes longer than 60 minutes across all three sensor wear locations: wrist, hip, and thigh. They were closely followed in performance by the decision tree models that included surface skin temperature as a predictive variable (`tree_full` and `tree_imp6`). On the other hand, the random forest model demonstrated excellent performance only on data from the wrist, delivering mediocre results on hip and thigh data. The `syed_CNN` and `tree_no_temp` yielded mediocre results across all wear-locations. When we shifted our focus to short non-wear episodes lasting less than 60 minutes, the limitations in the `cz_60` and `heu_alg` algorithms due to their built-in minimum episode durations of 60 and 20 minutes, respectively, became evident. Similarly, the deep learning model (`syed_CNN`) showed poor results, mainly attributable to very low sensitivity scores across all wear-locations that led to many episodes being misclassified as non-wear time. The random forest model's performance was also poor on the hip and only mediocre on the thigh and wrist. The decision tree models, both without temperature and with all predictors, showed mediocre performance as well. However, the decision tree model trained on the six most important predictors stood out as the best performer for short non-wear episodes on all wear-locations. This study also highlighted the value of incorporating surface skin temperature as a predictor to enhance the performance of non-wear time classification. Overall, these results provide valuable insights into the effectiveness of various methods for classifying non-wear episodes

in accelerometer data, emphasizing the potential of simple algorithms like `cz_60` and `heu_alg`, especially for longer non-wear episodes, and the benefit of including surface skin temperature as a predictive variable.

We discovered that most non-wear episodes in our ground truth datasets had a duration exceeding 60 minutes, with a noticeable peak around the 10-hour mark. This finding contrasts with previous research that typically reported shorter episodes as being more prevalent[113,129,149]. Our data prominently features children and physically active adolescents. While this demographic, known to spend less time in sedentary activities and frequently interrupt such periods[150], might remove wearables for sports or other activities, the extended non-wear durations, especially those of several hours, are unlikely to be mistaken for sedentary behavior. Such prolonged episodes more definitively signal non-wear rather than inactivity. This clearer distinction minimizes potential overlaps or ambiguities between non-wear and sedentary periods in our dataset. This also meant that the data favored the simple heuristic algorithms for classifying non-wear time, largely because the limitations imposed by minimum window lengths had a negligible impact on the proportion of non-wear time that was incorrectly classified. These algorithms achieved excellent precision scores, confirming that neither sedentary time nor sleep was misclassified as non-wear time. This is a significant finding, given that multiple previous studies have pointed out the complexities in making this very distinction[105,115,123,142,151,152].

Creating a model to classify non-wear time seems to be a relatively straightforward task. A primary reason is the distinct data patterns that wear and non-wear times produce. For instance, non-wear times should typically yield consistent zero or near-zero readings due to the absence of movement, while wear times elicit varied readings reflecting activity levels. This distinction implies that the decision boundary involved in this classification is likely close to linear. Given the binary nature of this classification task – the device is either worn or not – such clear data patterns could make more complex models unnecessary. In fact, using complex models, like the ones employed in the current study, can introduce the risk of overfitting. Overfitting tends to capture the random fluctuations or noise inherent in the training dataset, ultimately compromising the model's performance on unseen, out-of-sample data. In light of these observations, we hypothesize that a simpler, well-optimized logistic regression model could be just as effective, if not more so, than the complex models that was evaluated. The strength of logistic regression in this context lies in its ability to establish a separating linear hyperplane. This hyperplane may effectively differentiate between wear and non-wear times without the added intricacy of non-linear decision boundaries. Furthermore, complex models, while offering sophisticated decision-making capabilities, may be redundant for this task and could introduce unnecessary complications. This is especially evident when the objective is creating a universally applicable machine learning model suited for diverse populations and various wear locations. To optimize the performance of any chosen model, be it simple or complex, the quality and diversity of the training data are important. It is vital to source data that spans multiple wear locations and encompasses a variety of physical activity profiles. Training on such a comprehensive dataset ensures that the model is equipped to discern between wear and non-wear times across different scenarios and populations. While our discussion leans towards simpler models given the nature of the decision boundary, the emphasis remains on the crucial role of a rich and varied dataset

in ensuring model efficacy.

This study emphasizes the potential of the consecutive zeros algorithm as a leading approach for identifying non-wear episodes exceeding 60 minutes in children and adolescents. This method showed efficacy across various wear locations, such as the hip, thigh, and wrist. However, due to the distinctive behaviors and routines of these younger age groups, our results may not be seamlessly extended to older populations. In comparison, models like the `syed_CNN`, which detect accelerometer mounting and unmounting events, present certain complexities. The motions associated with putting on or taking off a device may not be universally consistent, as they could vary depending on age, dexterity, or other factors related to the population in question. Therefore, while the `syed_CNN` model employs a standardized approach, its results might still be influenced by age or other population-specific factors. Though our research provides valuable insights for younger demographics, it's essential to approach the `syed_CNN`'s results with an understanding of its potential limitations across varied age groups in accelerometer-based studies as well.

Incorporating surface skin temperature for the classification of non-wear time has been scarcely explored in the field of machine learning. One study did indicate that using acceleration data along with rate-of-change in surface skin temperature could create a robust decision tree model for detecting non-wear time[153]. This aligns with previous studies that have shown improved predictive performance when temperature data is included in heuristic algorithms[115,117]. Our own findings also corroborate this, as we observed that adding surface skin temperature as a feature enhanced the performance of the non-wear decision tree models. However, to accurately detect transitions between wear and non-wear periods, it may be beneficial to account for the slow response time of temperature sensors. Solely using temperature data can cause classification delays. By incorporating lagged and leading temperature features, we can better anticipate previous and upcoming temperature changes. Therefore, a combined approach using both temperature, with its nuanced features, and acceleration data is advised, as supported by Zhou and colleagues.[117]. During our study, we noted a 20-minute step response in the Axivity temperature sensor, which could be attributed to the design of the device's casing. The sensor's response time may also be influenced by the attachment method used. If more material is placed between the skin and the device, delays are likely to be amplified, suggesting that machine learning models should perhaps consider the type of sensor attachment in their feature set. Additionally, different brands of devices have been found to have varying optimal temperature thresholds, further complicating the issue. As noted by Duncan et al. and Zhou et al., algorithmic modifications are needed for devices to function optimally in different latitudes[115,117]. Therefore, the type of device and its attachment method can be critical components for improving the accuracy of non-wear time classification models.

Evaluating the performance of a machine learning model is a critical step in ensuring its reliability. Typically, researchers use a portion of the dataset they trained the model on, segmenting it for testing purposes. This is termed as "internal validation." While standard and useful, this method isn't without its shortcomings. To illustrate, the studies conducted by Syed et al. and Sundararajan et al. showcase impressive performance metrics, such as sensitivity, specificity, and accuracy in their task to classify non-wear time. Yet, a closer look reveals that these high-performance results are rooted in cross-validation

techniques that didn't incorporate external validation datasets[100,145]. This omission leads us to question the generalizability of their models. Can these models perform as well in real-world scenarios with diverse data as they did during testing? Models that are highly flexible and adaptive pose a risk, especially if they are not subjected to rigorous validation. There's the potential danger of them overfitting to the training data, meaning they might excel at recognizing patterns from the specific dataset they were trained on but falter when faced with new, unseen data. Such models might inadvertently learn the unique quirks and nuances of the training dataset rather than universally applicable patterns. This potential for overfitting becomes especially concerning when considering the methodology of Syed et al.'s model. Their approach, while promising, would greatly benefit from training on a dataset enriched with data from a broader spectrum of participants. Factors like age, lifestyle, or health can lead to subtle differences in how devices are mounted or unmounted, creating variations in signal shapes. By including a more diverse population in their training data, the model's ability to generalize and correctly interpret signals from different user groups would likely improve. Considering these insights, it's evident that the next wave of research in this domain should prioritize validating their models using independent, external datasets before they're widely accepted or published. We understand the challenges tied to accumulating extensive and diverse datasets, especially in niche fields. However, the promise of crafting a model that's both reliable and universally applicable makes this a worthwhile pursuit.

When analyzing acceleromtry data, the ideal scenario is to employ a single model that performs reliably across different wear locations and populations. To evaluate the generalizability and robustness of the developed decision tree models used in the present study, we included a dataset from wrist-worn devices for external validation. This ensures that the performance metrics of our decision tree models are not artificially inflated due to overfitting or lack of variance between the training and testing data. External validation involves testing a model with independently sourced datasets to confirm its performance. If a predictor set has been inaccurately selected or if model parameters have been overly tuned to characteristics inherent to the training data, such as technical or sampling bias, the model is likely to perform poorly during external validation[154]. The rationale for using external validation is compelling: although data from various sources might exhibit differences, they can still encapsulate crucial domain-specific information. A well-trained model that focuses on truly informative predictors should retain its performance when applied to new, previously unseen data. Therefore, the external validation in our study acts as a verification step, ensuring that our decision tree models that pass this criterion are not just robust but also likely to be interpretable within the domain[155]. While Syed et al.'s methodology for identifying non-wear time is innovative and logically coherent, we believe its performance may vary depending on the age of the population in the dataset and the employed attachment method since their approach focuses on identifying the specific shape of the acceleration signal at the start and end of a non-wear episode. In contrast, methods that simply identify non-wear time based on the absence of acceleration are less dependent on the characteristics of the population, since zero movement during non-wear is a universal trait. Our results support this idea. The `syed_CNN` model showed poor performance across all wear locations in our study.The diminished performance of the `syed_CNN` model might be attributed to differences in population characteristics, given that accelerometers in both the PHASAR study and the Tromsø Study[156] —which served as training data for `syed_CNN`—were affixed using elastic belts. Furthermore, the

reduced accuracy on wrist data was anticipated, as the signal shape generated during device removal from the wrist is distinct from that of the thigh or hip. The `syed_CNN` model was trained on an older population, aged between 40-84 years (mean = 62.74, SD = 10.25), whereas our study involved datasets of younger individuals aged 8.1-17.9 years (mean = 12.14, SD = 2.40) for hip and thigh data, and 14.5-16.4 years (mean = 15.4, SD = 0.37) for wrist data. Contrastingly, the `sunda_RF` model showed acceptable performance in identifying non-wear episodes shorter than 60 minutes on both the thigh and wrist data. This suggests that the model by Sundararajan et al. is less affected by dataset differences compared to the `syed_CNN` model. Another point worth noting is that the `syed_CNN` model was originally trained on data with a frequency of 100 Hz, while we applied it to data with frequencies of 50 Hz and 25 Hz. Although it's unclear whether this frequency difference impacted the model's performance, we believe that the 25 Hz data is sufficient for capturing true movement behavior, given that movement frequencies are generally below 5 Hz.

The robustness of this study is significantly enhanced by the use of external validation, which offers strong evidence of methodological generalizability. However, there are limitations to consider. One major issue is the absence of a universally accepted gold standard for ground truth datasets with non-wear periods in this research area. This lack of a benchmark makes it challenging to compare performance metrics across different studies. Despite this, our approach remains transparent since it relies on raw accelerometer data, and no part of our data collection or analysis process is proprietary. It's important to note that our findings are based on a study population consisting of children and adolescents. Consequently, the results may not be directly applicable to older age groups. Additionally, while we chose to develop decision tree models for their balance of complexity and interpretability, future research could explore the efficacy of other machine learning methods like logistic regression, gradient boosting, or support vector machines. These alternative algorithms may offer different insights or advantages that could improve upon our current model.

Paper III: Improving Sleep Quality Estimation in Children and Adolescents: A Comparative Study of Machine Learning and Deep Learning Techniques Utilizing Free-Living Accelerometer Data from Thigh-Worn Devices and EEG-Based Sleep Tracking

This segment of the thesis encompasses the methods, results, and discussion for Paper III which is in preparation for SLEEP (see Appendix III). Polysomnography, the premier method for sleep evaluation, is not always feasible for extensive research due to its high costs and impracticality. Wearable accelerometers present an affordable solution. While wrist and hip-worn devices dominate sleep studies, the potential of thigh-worn accelerometers remains largely untapped. The primary aim of this paper was to assess various machine and deep learning models designed to estimate in-bed and sleep time using raw data from a tri-axial, thigh-worn accelerometer. For a robust validation, the outcomes of these models were compared with results from the ZM, which served as our gold standard for sleep assessment in this study. Additionally, a secondary objective was to assess the efficacy of these models in determining key sleep quality metrics, including sleep period time (SPT), total sleep time (TST), sleep efficiency (SE), latency until persistent sleep (LPS), and wake after sleep onset (WASO).

Methods

Dataset and Participants

The current study uses data from the SCREENS trial[116,157], which took place from June 2019 to March 2021 in the Region of Southern Denmark. The trial aimed to evaluate the impact of limiting screen media usage among Danish families. We specifically analyzed data from children between the ages of 4 and 17, with a mean age of 9.1 years, who were part of the SCREENS cohort. Our primary sources of data were accelerometer readings from Axivity AX3 devices, and EEG-derived sleep states and sleep quality metrics from the ZM device. The children wore the Axivity AX3, a discreet 3-axis accelerometer, on their right thigh, halfway between the hip and knee, to capture their movement data.

The ZM, developed by General Sleep Corporation, was utilized to extract sleep state information. The ZM device integrates advanced EEG technology and signal processing

algorithms. Participants were instructed to attach three self-adhesive, disposable sensors outside their hairline when going to bed and to detach them upon waking up, ensuring consistent and clear EEG signal acquisition. Two key algorithms underpin the ZM: Z-ALG and Z-PLUS. Z-ALG is employed for sleep detection, proving ideal for single-channel EEG at-home monitoring[50]. Supplementing Z-ALG, Z-PLUS effectively discerns between sleep stages, aligning closely with expert-assessed PSG data[51]. However, in our study, we grouped all sleep stages (light sleep (N1 & N2), deep sleep (N3), and REM sleep) into a single "asleep" category. This approach streamlined the machine learning process as differentiating between sleep stages wasn't essential for our target sleep quality metrics.

Figure 11 outlines the selection criteria for children's recordings from the SCREENS study. Only recordings from ZM with complete accompanying accelerometer data and durations between 7 and 14 hours were considered. Any nights where ZM indicated sensor issues were discarded. This left us with a total of 585 nights from 151 children, averaging 3.87 nights per child (SD = 1.86). The age of these children averaged at 9.4 years with a standard deviation of 2.1 years. Across these recordings, ZM predictions spanned 696,779 epochs, each lasting 30 seconds with around 84% of the overall ZM recording time classified as sleep.

Lastly, we affirm that the SCREENS trials adhered to ethical guidelines, receiving approval from the Regional Scientific Committee of Southern Denmark. All data handling processes were in compliance with the General Data Protection Regulation (GDPR), ensuring the secure and ethical management of participant information.

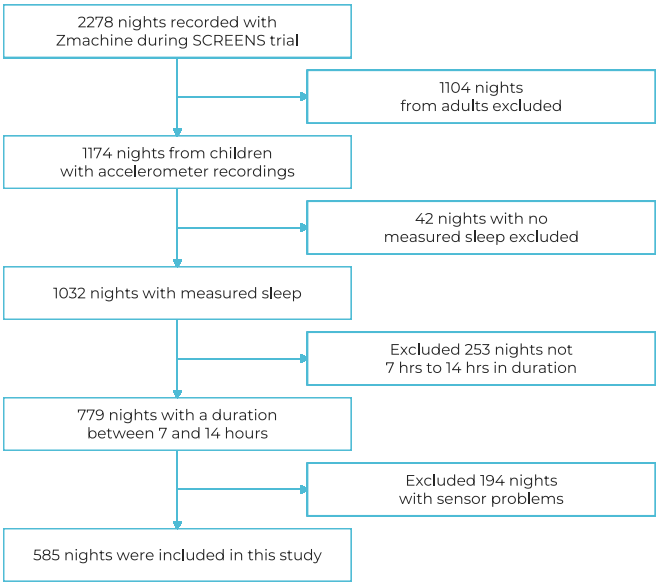

Figure 11: Flowchart depicting the selection process for eligible ZM recording nights included in the study.

Data preprocessing and Feature Extraction

In this study, we began by processing raw accelerometer data through a low-pass filtration step, utilizing a 4th order Butterworth filter with a 5 Hz cut-off frequency to remove high-frequency noise as described by Skotte and colleagues[63]. Non-wear data was identified and removed using the decision tree classifier, `tree_imp6`, as outlined in Paper II[141], and the remaining data was then resampled into 30-second epochs to match the granularity of the ZM recordings.We then conducted feature extraction, generating 64 features that offered a comprehensive characterization of the data. These features were derived from both accelerometer and temperature signals and included temporal elements, which utilized both lag and lead values to capture dynamic data trends. Additionally, we took inspiration from Walch et al.[158] to include sensor-independent features that encapsulate circadian rhythms, offering unique insights that are not directly discernible from sensor outputs (see Figure 12). Including both a cosine function and a linear function to represent circadian rhythm offers a nuanced depiction of time-based patterns in sleep data. The cosine function captures the inherent 24-hour rhythmicity of the circadian cycle, reflecting the natural ebb and flow of human behaviors and biological processes like melatonin secretion[159]. On the other hand, the linear function provides a continuous representation of time, illustrating the progression throughout the night and accounting for gradual changes in sleep propensity as one transitions from early night to early morning. By integrating both these features, the models can holistically understand the repetitive nature of circadian rhythms and the distinct characteristics of each night, thereby potentially enhancing its predictive accuracy. We further enriched the feature set by incorporating signal characteristics such as vector magnitude, mean crossing rate, skewness, and kurtosis for each of the x, y, and z dimensions. All features are summarized in Table 10. The ZM recordings and the corresponding accelerometer data were then merged. Any time overlap between these two sets of data was categorized as 'in-bed' time, while the remaining time was considered 'out-of-bed.' This process yielded a comprehensive dataset that provided a 24-hour view of each participant's activity and sleep patterns, with a target feature which consisted of three classes of interest; "out of bed awake", "in bed awake", and "in bed asleep".

Table 10: All extracted features grouped by category.

Feature Category	Count	Summary of Features
Misc Features	3	age, weekday, vector_magnitude
Inclination & Orientation	2	incl, theta
Signal Means	4	mean for x, y, and z
Signal SDs	8	SDs for temp, x, y, and z (30-sec and 15-min windows) and sd_max
Clock Proxies	2	clock_proxy_cos, clock_proxy_linear
Time-Dependent	36	1-, 5-, and 10-minute lag and lead features
Crossing Rates	3	mean crossing rate for x, y, and z
Signal Distributions	6	kurtosis and skewness for x, y, and z
Total	**64**	

Upon examining the raw ZM predictions, we observed that the device appeared to overestimate the number of awakenings among the children studied. Although the ZM software addresses many of these awakenings by counting only three consecutive awake

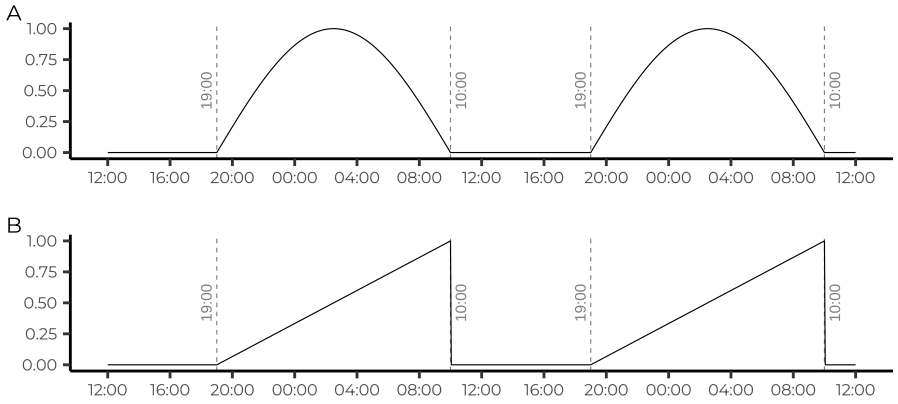

Figure 12: Sensor-independent features of circadian rhythms across two consecutive nights. A) cosinus feature, B) linear feature.

epochs towards wake time, this approach renders the raw predictions less suitable as training data for machine learning algorithms. In fact, many of these awakenings, labeled by the ZM, would be more aptly described as arousals rather than actual awakenings. Separately, the ZM device's sleep efficiency rating for our sample was 83%, which is below recognized standards. An efficiency of 85% is considered good, and over 90% is seen as ideal. This contrasts with prior research on similar child cohorts that reported a sleep efficiency of 88.3%[160]. Recommendations from an expert panel by the National Sleep Foundation emphasized that fewer than 2 awakenings lasting more than 5 minutes each night qualify as good sleep across all age groups[161]. Additionally, it's widely recognized that children typically experience between five to eight sleep cycles every night, with awakenings most likely occurring at the conclusion of each cycle[162]. However, definitions of a "waking bout" vary across studies. Some demand at least 5 continuous minutes of wakefulness for it to be counted as one bout, while others find a 1-minute duration adequate. Of particular note, the vast majority short arousal epochs labeled as awake by the ZM did not show any relevant responses in the accelerometer signal as inferred by visual examination. This misalignment might distort underlying patterns for machine learning algorithms. While this might not be outright mislabeling, categorizing all such epochs as true awakenings would introduce noise, jeopardizing model accuracy. In light of these observations, we opted to process both the raw ZM output and versions with 5-minute and 10-minute median filtering for our model training and evaluation. This approach minimized noise and offered an awakening count more aligned with typical patterns in children's sleep (see Table 12 and Figure 13 for details).

Algorithms and Modelling Strategies

The task of analyzing sleep patterns based on accelerometer data has an inherent hierarchical nature. Broadly, the goal is to distinguish between the states of being 'out-of-bed' versus 'in-bed'. Once the 'in-bed' state is determined, the finer classification into 'awake'

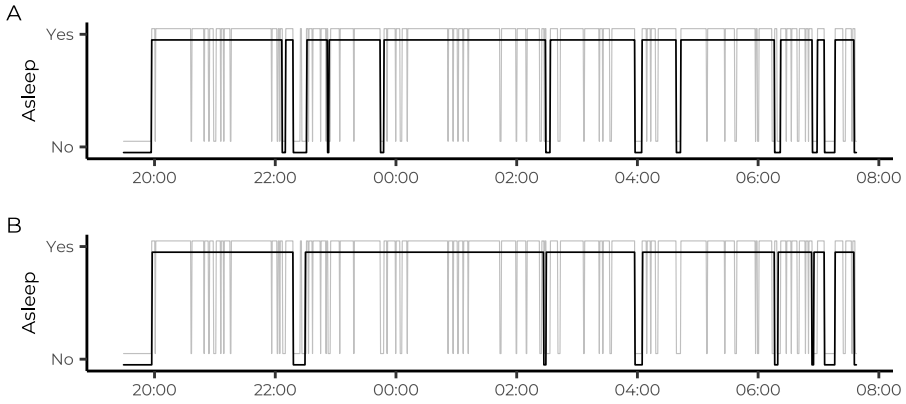

Figure 13: The difference in number of awakenings between the raw ZM predictions vs. 5-minute, and 10-minute median filtered predictions for a random night (boy, 9 years). Grey line is the raw predictions, black line is the median filtered predictions. A: 5-minute median filter on raw ZM predictions, B: 10-minute median filter on raw ZM predictions.

or 'asleep' states follows. This hierarchical structure of the problem informs the first of our modeling strategies.

Our first model strategy leveraged this hierarchical structure by deploying a sequence of two models, each functioning as a binary classifier. This method simplified the prediction task by breaking down the multiclass problem into two binary phases: first identifying 'in-bed' periods, followed by determining 'sleep' periods. The output from the first binary classifiers in sequence, which estimated in-bed time, underwent a 5-minute median filter to eliminate blips of in-bed time. This step allowed us to define a singular continuous interval recognized as the SPT, representing the total duration spent in bed attempting to sleep. This SPT subsequently served as input for the next stage of binary classifiers, which predicted sleep periods within the SPT.

Four machine learning algorithms were employed in pairs in this sequential strategy:

1. Logistic Regression: Logistic regression served as a simple and fast baseline model. However, due to its linear nature, it may struggle with capturing complex relationships and non-linear patterns present in the accelerometer data.

2. Decision Tree: Decision trees are capable of handling non-linear patterns and are easily interpretable. However, they are prone to overfitting, particularly when dealing with complex patterns that require simultaneous consideration of multiple features. To combat this, we used a maximum tree depth of 8.

3. Single-layer Feed-forward Neural Network (MLP): Also know as a multi-layer perceptron, MLPs can effectively capture non-linear relationships, even with their relatively simple structure. However, they tend to be more challenging to interpret compared to simpler models. Additionally, careful tuning of the network's architecture and training process is required to mitigate the risk of overfitting.

4. XGBoost: XGBoost is a powerful algorithm known for its ability to provide highly accurate predictions and handle complex, non-linear patterns in the data. It also incorporates built-in methods to prevent overfitting. However, training XGBoost models can be computationally intensive, and interpreting the predictions it generates can pose challenges.

In contrast to the hierarchical approach, we also experimented with a standalone strategy using a bidirectional Long Short-Term Memory (biLSTM) neural network[163] as a multiclass classifier. Opting not to use the biLSTM in the sequential approach was about strategy diversification rather than concerns over complexity or interpretability. We wanted to compare a sequence of traditional machine learning models using an hierarchical approach with a standalone deep learning approach using the biLSTM. This distinction would underscore the advantages and limitations of each approach. Our choice to employ the biLSTM arose from its inherent capability to capture temporal sequences in accelerometer data, especially when distinguishing between the three states: 'out-of-bed', 'in-bed-awake', and 'in-bed-asleep'. Each of these states doesn't exist in isolation; they transition from one to another in patterns that are critical for accurate predictions. For instance, understanding the sequential pattern leading from 'in-bed-awake' to 'in-bed-asleep' can provide vital context, and the biLSTM is adept at capturing such temporal dependencies. Specifically designed for these three classes, the model comprised of four layers and 128 hidden units per layer, balancing complexity and efficiency. The bidirectional setup enhanced data comprehension while mitigating overfitting risks. The model processed 10-minute tensor sequences with a one-epoch step size. Previous studies, such as Sano et al.[164] and Chen et al.[165], have evidenced the effectiveness of LSTM models aiding in detecting sleep from accelerometer data.

Model Training

We trained a total of four pairs of models in sequence, with each pair distinguishing between in-bed/out-of-bed and asleep/awake classes, respectively. The dataset was randomly split into a training and a testing set, each containing approximately half of the subjects. We made sure that data from the same night was not distributed across both sets. This approach was adopted to ensure that the model could effectively generalize to unfamiliar data, rather than overfitting to specific participant data similar to the data partitioning in Paper II. To tune the hyperparameters of our models, we used a specific set of hyperparameters for each type of machine learning algorithm. For the Decision Tree, we tuned the cost complexity, tree depth, and minimum number of samples required at a leaf node. The decision tree model was set up using the rpart[148] engine, with tree depth ranging from 3 to 7. For Logistic Regression, implemented using the glmnet[166] engine, we tuned the penalty and mixture hyperparameters controlling regularization. The MLP was implemented with a single-layer feed-forward architecture using the nnet[167] engine, with the maximum number of allowable weights (MaxNWts) set to 7000 as a form of regularization to constraint the model from becoming too large and possibly overfitting the training data. The hyperparameters we tuned for this model were the number of hidden units, the penalty, and the number of epochs. The range for the number of hidden units was between 3 and 27. Lastly, the XGBoost model was configured with the xgboost[168] engine. The hyperparameters subjected to tuning included tree depth,

learning rate, loss reduction, minimum number of samples required at a leaf node, sample size, and number of trees. For this algorithm, the number of trees was specifically tuned within a range of 200 to 800 to constraint training time. See Table 11 for an overview of the model-specific hyperparameters. We optimized all model hyperparameters using a 10-fold Monte Carlo cross-validation technique, which involves randomly partitioning the dataset into training and validation sets across multiple iterations. The optimization of hyperparameters was carried out via a grid search paired with latin hypercube sampling. By defining specific ranges for each hyperparameter, latin hypercube sampling systematically divides these ranges into segments and draws a random value from each, ensuring a well-distributed set of hyperparameter combinations. The grid search then methodically assessed various combinations of these hyperparameters to determine which combination produced the best model performance.

Table 11: Details of the hyperparameters tuned for each machine learning model, their descriptions, and the specific range from which values were sampled during grid search optimization.

Model	Hyperparameter	Description	Range
Logistic Regression	Regularization Strength	Controls magnitude of regularization penalty	[-10, 0] (log scale)
	Lasso Proportion	Mix between Ridge (0) and Lasso (1) regularization	[0, 1]
Decision Tree	Cost Complexity	Controls the trade-off between tree's depth and fit	[-10, -1] (log scale)
	Tree Depth	Maximum depth of the tree	[3, 7]
	Minimal Node Size	Minimum number of samples required to split a node	[2, 40]
Multi-layer Perceptron	Hidden Units	Number of units in the hidden layer(s)	[1, 10]
	Regularization Strength	Controls magnitude of regularization penalty	[-10, 0] (log scale)
	Number of Epochs	Number of complete passes through the dataset	[10, 1000]
XGBoost	Learning Rate	Step size shrinkage to prevent overfitting	[-10, -1] (log scale)
	Minimum Loss Reduction	Minimum loss reduction required for partition	[-10, 1.5] (log scale)
	Minimal Node Size	Minimum number of samples required to split a node	[2, 40]
	Observations Sampled	Proportion of samples used per iteration	[0.5, 1]
	Number of Trees	Total number of trees to train	[200, 800]

After identifying the best-performing hyperparameters, we proceeded to fit the models to the full training dataset. This approach allowed us to use all available data for model parameter estimation, thereby maximizing performance. To tackle the imbalance in the extracted in-bed time from our sequential modelling strategy, where the 'awake in-bed' class made up only about 15% of the training data for the sleep/wake classifiers, we employed the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by generating synthetic samples in the feature space to balance out the classes. Specifically, it creates synthetic observations by randomly selecting a minority class instance and its nearest neighbors, then producing a new instance that is a blend of the two. This method, as outlined by Chawla et al.[169], mitigates biases during model training that arise when models are skewed towards the majority class. Using the themis R package[170], we implemented SMOTE to achieve a balanced distribution of training samples for both classes. The optimization was driven by the F1 score as a metric since it harmonizes precision and recall, rendering it more resilient to class imbalance.

In parallel to these sequential models, we trained the biLSTM model to classify three distinct classes: "out-of-bed awake", "in-bed awake", and "in-bed asleep". The data for this model was divided into training, validation, and test sets, adhering to a 50/25/25 split ratio. Again, caution was exercised to avoid having data from the same night across different sets. For efficient and adaptive learning, the Adam optimizer was used during the training process. Given that we were dealing with a multiclass classification task with mutually exclusive classes, the cross-entropy loss function was employed. At the output layer, a softmax activation function was applied to obtain a probability distribution over the classes. We employed early stopping with a patience of 3 epochs, ceasing training if no improvement in the validation loss was observed over three consecutive epochs.

Model Validation

In the current study, we utilized standard evaluation metrics derived from confusion matrices to assess the performance of each model on an epoch-to-epoch basis. These include

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$precision = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$F_1 = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity}$$

where NPV is negative predictive value, F_1 is the F1 score, TP is true positives, FP is false positives, TN is true negatives, and FN is false negatives.

In our sequential modelling strategy, the models in the first stage that carried out the binary classification task of distinguishing between in-bed and out-of-bed states was evaluated using the F1-score, accuracy, sensitivity, specificity, and precision. In the second stage of the sequential modelling strategy, models responsible for distinguishing between 'asleep' and 'awake' states were evaluated using the same metrics, with the addition of the negative predictive rate. Given the class imbalance, the F1 score was calculated as an unweighted macro-average, allowing the metric to represent predictions for both classes effectively. We also scrutinized a multiclass biLSTM classifier using the same metrics, interpreting its multiclass output as two separate binary classifications: out-of-bed versus all other states, and in-bed-awake versus in-bed-asleep. Moreover, to give a comprehensive view of model performance, we present confusion matrices for the entire dataset, covering both in-bed and out-of-bed data. These matrices report relative counts, column percentages for accurate predictions of the true class, and row percentages for correctly classified predictions. Both in-bed/out-of-bed and awake/asleep classification tasks were treated as binary, designating 'in-bed' and 'asleep' as positive labels and 'out-of-bed' and 'awake' as negative labels, in line with prior studies[72,171].

To evaluate how well our models performed in generating sleep quality metrics, we employed Bland-Altman plots and Pearson correlations. Specifically, the Bland-Altman approach was used to estimate the level of agreement between two different measurement techniques. Given the nature of our dataset, which contains multiple observations per subject but not necessarily equal number of observations, we used a bootstrap procedure to account for this extra variability. We initially calculated the mean difference or bias, and then determined the limits of agreement (LOA) as the bias ± 1.96 times the standard deviation of these differences. Given the possibility of non-normal distribution and skewness in our data, we opted for a bias-corrected and accelerated bootstrap method[172]. This allowed for more accurate estimation, taking into account intra-subject variability. Using 10,000 bootstrap replicates, we estimated the 95% confidence intervals for both the bias and LOA, thereby ensuring robust measurements. The sleep quality metrics conformed to ZM definitions and included the following:

1. Sleep Period Time (SPT) - This refers to the total duration of time in bed with the intention to sleep, which is defined as the time from the start to the end of the ZM recording.
2. Total Sleep Time (TST) - This is the time spent asleep within the SPT.
3. Sleep Efficiency (SE) - This is the ratio between TST and SPT, representing the proportion of the sleep period that was actually spent asleep.
4. Latency Until Persistent Sleep (LPS) - This metric represents the time it takes to transition from wakefulness to sustained sleep. It is calculated as the time from the beginning of the ZM recording until the first period when 10 out of 12 minutes are scored as sleep.
5. Wake After Sleep Onset (WASO) - This refers to the time spent awake after initially falling asleep and before the final awakening. In our analysis, a period is counted as 'awake' only if it consists of 3 or more contiguous 30-second epochs which is also how the ZM summarizes WASO.

The technical frameworks used for model development and analyses were R version 4.3.0[173] along with the Tidymodels[147] and Tidyverse[174] package suites. For the biLSTM

model, we used Python version 3.10.6[175] and PyTorch[176].

Results

As indicated in Table 12, the application of 5-minute and 10-minute median filters led to modifications in the sleep quality metrics derived from ZM predictions. SPT remained consistent between raw and filtered data sets, with a mean duration of 9.2 ± 2.1 hours, which aligns with the length of the ZM recording. TST and SE increased in the filtered data, implying the filters designate some wakefulness as sleep. Specifically, the mean TST rose from 7.7 ± 1.9 hours in the raw data to 8.1 ± 2.0 hours with a 5-minute filter and to 8.2 ± 2.1 hours with a 10-minute filter. Similarly, SE increased from an initial mean of 82.6 ± 12.0% to 86.4 ± 12.7% and 87.5 ± 12.9% for the 5-minute and 10-minute filters, respectively. Furthermore, the LPS also saw an increase, implying that the filters are removing brief asleep periods at the onset of sleep, thereby lengthening the time it takes to achieve persistent sleep (i.e., 10 out 12 minutes classified as asleep). On the other hand, the WASO metric decreased from a raw average of 39.0 ± 33.6 minutes to 30.6 ± 46.8 minutes and 22.3 ± 55.4 minutes in the 5-minute and 10-minute filtered data, respectively. Notably, the application of these filters also led to a significant reduction in the average number of awakenings per night. In the unfiltered data, the mean number of awakenings (or arousals) stood at 34.46 ± 11.33, which dropped to 4.43 ± 3.26 and 1.95 ± 2.01 in the 5-minute and 10-minute filtered datasets, respectively.

Table 12: Overview of characteristics of the ZM sleep quality summaries per night (585 nights from 151 children). Values are represented as mean (SD). Hrs: hours, min: minutes.

	SPT (hrs)	TST (hrs)	SE (%)	LPS (min)	WASO (min)	Awakenings (N)
Raw ZM Predictions	9.2 (2.1)	7.7 (1.9)	82.6 (12)	34.5 (27.9)	39 (33.6)	34.5 (11.3)
5-Min Median	9.2 (2.1)	8.1 (2)	86.4 (12.7)	36.3 (39.8)	30.6 (46.8)	4.4 (3.3)
10-Min Median	9.2 (2.1)	8.2 (2.1)	87.5 (12.9)	38 (48.7)	22.3 (55.4)	1.9 (2)

Performance on Epoch-to-Epoch Basis

As delineated in Table 13, the epoch-to-epoch evaluation for predicting in-bed time shows virtually identical performance across the various model types. The F1 score fluctuates slightly, ranging from 94.4% in the Decision Tree model to 95.4% in the XGBoost model. Likewise, accuracy varies minimally from 95.3% for the Decision Tree model to 96.1% for the XGBoost model. Other metrics such as sensitivity, precision, and specificity also exhibit uniform performance across the different models. While the XGBoost model does exhibit the highest performance with an F1 score of 95.4% and an accuracy of 96.1%, it only marginally surpasses the other models in these metrics.

Table 13: Performance metrics of the classification of in-bed/out-of-bed time of the included models.

	F1 Score (%)	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)
Decision Tree	94.4	95.3	93.1	95.6	96.9
Logistic Regression	95.0	95.7	95.0	94.9	96.3
Feed-Forward Neural Net	95.0	95.8	95.1	95.0	96.3
XGBoost	95.4	96.1	95.8	94.9	96.2
biLSTM	95.2	95.3	95.3	95.1	95.3

Table 14 illustrates the performance metrics for differentiating sleep/wake of all included models. In raw ZM predictions, the XGBoost model stood out with an F1 score of 76.2% and a precision of 92.8%. However, the biLSTM model struggled in this category, particularly with specificity, which was the lowest at 26.9%, even though its sensitivity was notably high at 98.1%. When a 5-minute median filter was applied, the XGBoost model further improved with an F1 score of 79.2%, and an increased NPV of 74.0%. The Decision Tree, during this phase, achieved an F1 score of 75.5% but a decreased specificity of 59%. The Neural Network's performance remained consistent, with an F1 score around 71.7% and precision of around 95.8%. The application of a 10-minute median filter saw the XGBoost's performance peak with an F1 score of 80.9% and a precision of 94.9%. In contrast, the biLSTM improved slightly with an F1 score of 70.9% but still lagged in specificity at 42.4%. Overall, while models like XGBoost seemed to demonstrate the most potential, the consistent challenge across models remained in achieving high specificity values.

Table 14: Performance metrics of the sleep/wake classification of the included models.

	F1 Score (%)	Precision (%)	NPV (%)	Sensitivity (%)	Specificity (%)
Raw ZM Predictions					
Decision Tree	72.9	93.2	48.4	86.3	67.1
Logistic Regression	71.0	93.7	43.9	82.7	70.9
Neural Network	71.8	93.8	45.1	83.6	70.8
XGBoost	76.2	92.8	58.0	91.3	62.8
biLSTM	65.6	86.1	75.0	98.1	26.9
5-Min Median					
Decision Tree	75.5	94.2	55.5	93.4	59.0
Logistic Regression	68.3	95.8	36.0	81.4	74.8
Neural Network	71.7	95.8	41.6	85.6	73.1
XGBoost	79.2	93.9	74.0	97.3	54.7
biLSTM	70.3	87.3	81.8	98.4	33.9
10-Min Median					
Decision Tree	76.3	94.7	58.1	94.9	57.5
Logistic Regression	68.0	96.5	34.3	81.9	76.4
Neural Network	71.0	96.1	39.5	86.5	71.4
XGBoost	80.9	94.9	75.8	97.7	57.6
biLSTM	70.9	89.2	61.0	94.6	42.4

Figure 14 and Figure 15 presents a comprehensive set of confusion matrices generated from data that includes both out-of-bed and in-bed periods. These matrices offer insights into the epoch-to-epoch performance of all sequential models when differentiating between 'awake' and 'asleep' states, irrespective of whether the subject is in bed or out of

bed. However, it's crucial to acknowledge that the sequential models, owing to their binary nature, are not equipped to directly classify the 'in-bed-awake' state. In contrast, the biLSTM model, which does identify the 'in-bed-awake' state as a separate class, seems to be less successful in classifying this specific state.

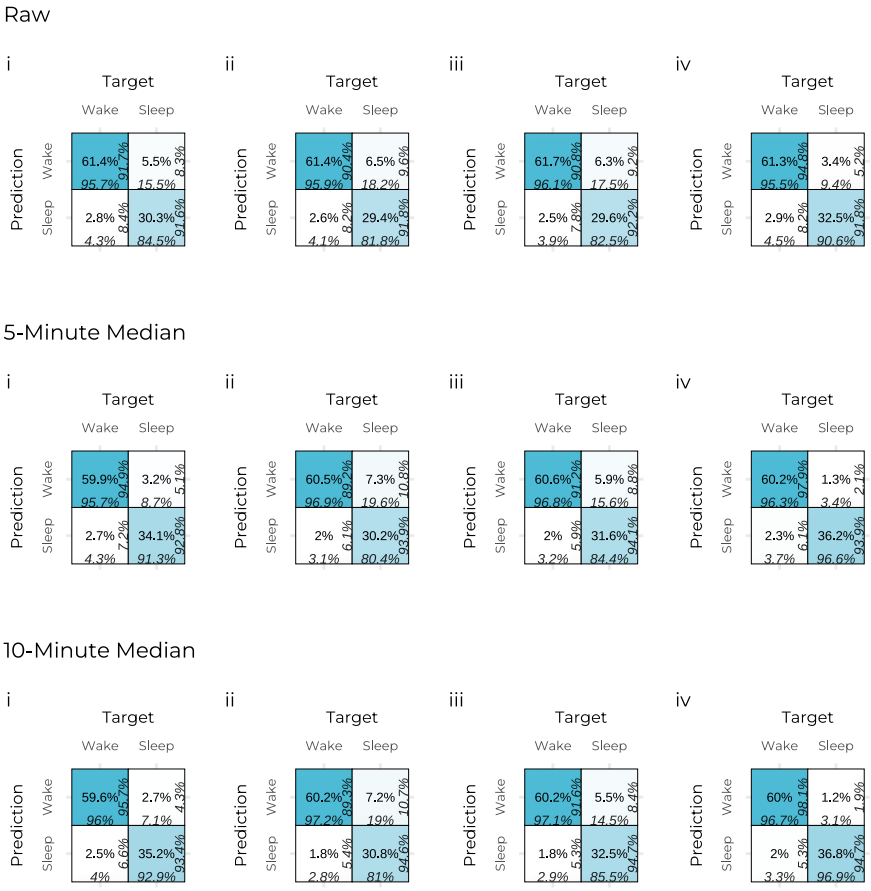

Figure 14: Confusion matrices for the binary predictions. The middle of each tile is the normalized count (overall percentage). The bottom number of each tile is the column percentage and the right side of each tile is the row percentage. i) decision tree, ii) logistic regression, iii) MLP, iv) XGBoost

Evaluation of Sleep Quality Metrics

Table 15 presents a comparative analysis of the included models used to predict various sleep quality metrics (SPT, TST, SE, LPS, WASO) using the 5-minute median filtered ZM predictions.

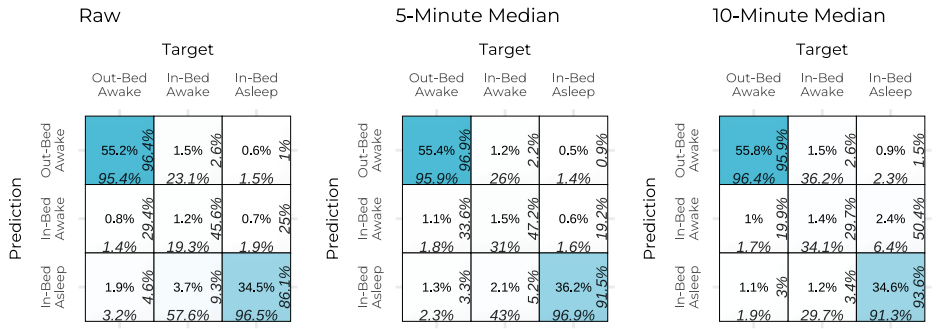

Figure 15: Confusion matrices for the biLSTM predictions. The middle of each tile is the normalized count (overall percentage). The bottom number of each tile is the column percentage and the right side of each tile is the row percentage.

Against the sleep quality metrics derived from the ZM raw data, the biLSTM model showed a -36.7-minute bias for SPT, while XGboost stood out for its minimal bias and correlation score of 0.56. Across metrics, varying biases emerged; for instance, in SE, the MLP and biLSTM displayed contrasting biases. For LPS and WASO, biLSTM generally showed strong biases, with overall correlations being poor.

Using the sleep quality metrics of the 5-minute median filtered ZM data, biLSTM's bias for SPT remained large. XGboost consistently showed the best correlations and smallest biases. The correlations for SE, LPS, and WASO, however, remained weak across most models whith XGboost again performing best.

Lastly, against the sleep quality metrics of 10-minute median filtered ZM data, biLSTM's bias persisted, especially in SPT, whereas XGboost and the Decision Tree exhibited moderate correlations. Notably, Logistic Regression and the MLP displayed pronounced biases, especially in TST and SE. Overall, while biases varied across models and metrics, again, XGboost consistently showed the highest correlations with concurrent small biases, indicating its stronger predictive performance relative to other models.

Overall, the decision tree model consistently underestimated SPT, TST, and SE, and overestimated LPS and WASO in comparison to ZM. The logistic regression model had similar trends, with more pronounced underestimation in TST and overestimation in LPS. The MLP also exhibited similar bias as the decision tree and the logistic regression models, but with a higher overestimation in WASO. On the other hand, the XGBoost model showed least bias among all, especially in its 5-minute median predictions. The biLSTM was the only model to overestimate TST, and as a consequence overestimate SE and underestimate LPS and WASO. This was true across raw and 5-minute filtered data but this trend did not apply to the 10-minute filtering. Considering LOA, the decision tree had higher variability in the differences across different sleep quality metrics and filterings, particularly for LPS and WASO, which indicates lower agreement with ZM. Other models had comparable LOA but with notable exceptions. For example, TST LOA for the logistic regression model was particularly wide in the 5-minute median predictions. Correlation-wise, the pearson coefficient, revealed that the XGBoost model consistently

had the highest correlation with ZM across all sleep quality metrics and filtering methods Notably, the XGBoost's 5-minute median predictions showed the strongest correlation (0.66) for TST among all models and filterings.

Table 15: Summary of bias, limits of agreement, and Pearson correlation for the included sleep quality metrics (SPT, TST, SE, LPS, WASO) across all included machine learning and deep learning models (decision tree, logistic regression, MLP, XGBoost, and biLSTM) on raw ZM predictions, 5-minute and 10-minute median predictions. Each value is provided with its 95% confidence interval.

	Bias (95% CI)	lower LOA (95% CI)	upper LOA (95% CI)	Pearson, r (95% CI)
Raw ZM Predictions - SPT (min)				
Decision Tree	-21.6 (-25.6;-17.6)	-117.5 (-125.6;-110.7)	74.2 (63.9;85.9)	0.54 (0.48;0.6)
Logistic Regression	-4 (-8.3;0.7)	-113.5 (-122.7;-106.1)	105.5 (95;118.7)	0.37 (0.29;0.43)
Feed-Forward Neural Net	-3.9 (-8.1;0.9)	-112.7 (-122;-105.2)	104.9 (94.1;118.4)	0.38 (0.3;0.44)
XGboost	0.2 (-3.7;4.5)	-97.4 (-106.2;-90.3)	97.8 (86.6;111)	0.56 (0.5;0.61)
biLSTM	-36.7 (-42.6;-30.3)	-141.4 (-153.2;-132)	68 (54.5;85.5)	0.5 (0.4;0.58)
Raw ZM Predictions - TST (min)				
Decision Tree	-148 (-153.8;-142.4)	-283 (-295.5;-272.6)	-13.1 (-22.8;-1)	0.3 (0.22;0.37)
Logistic Regression	-139.2 (-145.7;-132.8)	-291.6 (-306.1;-279.2)	13.1 (3.8;23.5)	0.12 (0.04;0.2)
Feed-Forward Neural Net	-154 (-159.9;-148)	-297 (-308.6;-287)	-10.9 (-20;-0.5)	0.25 (0.17;0.32)
XGboost	-66 (-70.8;-61.4)	-178.1 (-187.9;-169.6)	46.1 (38.9;54.5)	0.47 (0.4;0.53)
biLSTM	39 (33.3;44.9)	-60.1 (-72.9;-51.1)	138 (126;152)	0.53 (0.44;0.61)
Raw ZM Predictions - SE (%)				
Decision Tree	-22.7 (-23.7;-21.8)	-45.5 (-47.5;-43.8)	0 (-1.6;1.9)	0.17 (0.09;0.24)
Logistic Regression	-23.1 (-24;-22.1)	-45.6 (-47.4;-44)	-0.6 (-2;1)	0.18 (0.1;0.26)
Feed-Forward Neural Net	-25.6 (-26.5;-24.7)	-48.2 (-50;-46.6)	-3 (-4.5;-1.2)	0.23 (0.15;0.31)
XGboost	-11.1 (-11.8;-10.4)	-28.8 (-30.2;-27.5)	6.5 (5.5;7.7)	0.37 (0.29;0.44)
biLSTM	12.6 (11.8;13.3)	0 (-1.6;1.1)	25.2 (23.6;27.2)	0.07 (-0.05;0.18)
Raw ZM Predictions - LPS (min)				
Decision Tree	28.9 (24.5;33.2)	-76 (-87.6;-69.8)	133.8 (124.6;144.7)	0.13 (0.05;0.21)
Logistic Regression	47.5 (43.6;51.4)	-46.2 (-57;-38.6)	141.2 (131;154.5)	0.1 (0.01;0.18)
Feed-Forward Neural Net	34.3 (30.2;38.6)	-67.7 (-80.1;-60.5)	136.4 (126.4;149.3)	0.11 (0.03;0.19)
XGboost	34.5 (30.6;38.5)	-62.4 (-75.8;-55.2)	131.3 (121.1;143.9)	0.2 (0.12;0.28)
biLSTM	-17.6 (-24.1;-11.3)	-127.2 (-177.4;-97.4)	92.1 (63.4;143.8)	0.05 (-0.06;0.17)
Raw ZM Predictions - WASO (min)				
Decision Tree	46.1 (43;49.4)	-33.2 (-43.4;-26.2)	125.4 (117.7;138.8)	0.29 (0.22;0.37)
Logistic Regression	48.7 (45.3;52.1)	-34.7 (-46.2;-28)	132.1 (124.6;147.3)	0.25 (0.17;0.33)
Feed-Forward Neural Net	58.7 (55.4;62.1)	-23.8 (-33.9;-17.5)	141.2 (133.9;155.6)	0.33 (0.26;0.4)
XGboost	18.4 (15.6;21.2)	-50.2 (-62.7;-43.1)	86.9 (79.8;104.2)	0.36 (0.28;0.43)
biLSTM	-15.9 (-21;-8.9)	-116.1 (-158.9;-95.4)	84.3 (58.9;138.2)	0.04 (-0.07;0.16)
5-Min Median - SPT (min)				
Decision Tree	-21.6 (-25.6;-17.6)	-117.5 (-125.6;-110.7)	74.2 (63.9;85.9)	0.54 (0.48;0.6)
Logistic Regression	-3.7 (-8;1)	-112.2 (-120.9;-105.2)	104.8 (94;117.4)	0.38 (0.3;0.44)
Feed-Forward Neural Net	-3.9 (-8.1;0.9)	-112.7 (-122;-105.2)	104.9 (94.1;118.4)	0.38 (0.3;0.44)
XGboost	0.2 (-3.7;4.5)	-97.4 (-106.2;-90.3)	97.8 (86.6;111)	0.56 (0.5;0.61)
biLSTM	-36.1 (-41.7;-30)	-136.1 (-146.3;-126.9)	64 (51.1;78.6)	0.54 (0.45;0.62)
5-Min Median - TST (min)				
Decision Tree	-50.5 (-55.2;-46)	-161.4 (-175.8;-151.3)	60.4 (51.5;71.7)	0.48 (0.42;0.54)
Logistic Regression	-139.7 (-146.9;-133)	-305.6 (-323.6;-291.8)	26.2 (16.1;38.6)	0.09 (0.01;0.17)
Feed-Forward Neural Net	-126.5 (-132.8;-120.3)	-276.8 (-291.3;-264.7)	23.9 (14.8;33.9)	0.25 (0.17;0.32)
XGboost	-7 (-10.8;-3.3)	-95.5 (-105.2;-88)	81.4 (72.4;92.5)	0.66 (0.61;0.7)
biLSTM	12.8 (7.4;18.3)	-80.1 (-89.8;-72.3)	105.8 (94.3;118.8)	0.63 (0.55;0.69)
5-Min Median - SE (%)				
Decision Tree	-5.5 (-6.3;-4.7)	-23.9 (-26.4;-22.2)	12.9 (11.6;14.6)	0.22 (0.14;0.29)
Logistic Regression	-23.2 (-24.3;-22.2)	-48.1 (-50.9;-46.1)	1.7 (0.1;3.8)	0.13 (0.05;0.21)
Feed-Forward Neural Net	-20.9 (-21.9;-19.9)	-44.3 (-46.3;-42.5)	2.5 (1.1;4)	0.21 (0.13;0.29)
XGboost	-1.1 (-1.7;-0.5)	-15.6 (-17;-14.4)	13.3 (12.2;14.7)	0.44 (0.38;0.51)

biLSTM	8 (7.2;8.8)	-5.1 (-6.8;-3.8)	21.1 (19.5;23.1)	0.16 (0.04;0.27)
5-Min Median - LPS (min)				
Decision Tree	24.6 (19.7;29.1)	-88.8 (-115;-77.3)	138 (126.2;156.7)	0.06 (-0.02;0.14)
Logistic Regression	58.1 (53.4;62.6)	-52.3 (-75;-40.1)	168.6 (155.9;187.7)	0.05 (-0.03;0.13)
Feed-Forward Neural Net	35.3 (30.7;39.8)	-75.8 (-102.3;-63.4)	146.5 (134.4;166.9)	0.07 (-0.01;0.15)
XGboost	28.5 (23.9;32.6)	-76.4 (-104.2;-63.3)	133.4 (120.4;154.2)	0.12 (0.04;0.2)
biLSTM	-15.7 (-25.9;-7.5)	-169 (-230.7;-127.9)	137.6 (101.1;184.9)	0.09 (-0.02;0.2)
5-Min Median - WASO (min)				
Decision Tree	9.9 (6.5;14)	-79.4 (-109;-63.1)	99.2 (80;136.1)	0.15 (0.07;0.22)
Logistic Regression	45.4 (41.7;49.7)	-50.7 (-74.4;-38.4)	141.5 (126.8;173)	0.19 (0.11;0.27)
Feed-Forward Neural Net	45 (41.2;49.2)	-51.8 (-76.4;-39.1)	141.7 (125.8;174.1)	0.21 (0.14;0.29)
XGboost	-0.9 (-3.9;3)	-83.4 (-113.1;-66)	81.7 (62;119.6)	0.26 (0.18;0.33)
biLSTM	-3 (-9.9;7.7)	-144.1 (-197.2;-107.2)	138.1 (90.8;211.4)	0.02 (-0.1;0.13)
10-Min Median - SPT (min)				
Decision Tree	-21.8 (-25.7;-17.8)	-117.3 (-125.2;-110.4)	73.7 (63.4;85.4)	0.54 (0.48;0.6)
Logistic Regression	-4.2 (-8.6;0.5)	-113.4 (-122.4;-106)	105 (94.2;118)	0.37 (0.3;0.44)
Feed-Forward Neural Net	-4.1 (-8.5;0.6)	-112.6 (-121.7;-105)	104.5 (93.5;117.6)	0.38 (0.31;0.45)
XGboost	0.2 (-3.8;4.4)	-97.4 (-106.1;-90)	97.9 (86.7;111.1)	0.56 (0.5;0.61)
biLSTM	-83.7 (-90.7;-76.1)	-207.4 (-221.3;-195.2)	40 (27.7;57.1)	0.3 (0.19;0.4)
10-Min Median - TST (min)				
Decision Tree	-31.5 (-35.7;-27.4)	-129.9 (-140.9;-121.8)	67 (58.3;77.7)	0.56 (0.5;0.61)
Logistic Regression	-130.9 (-138;-124.2)	-295.1 (-311.8;-281.4)	33.2 (23.3;45.1)	0.09 (0.01;0.17)
Feed-Forward Neural Net	-116.3 (-122.9;-110.3)	-266.2 (-280.4;-254.2)	33.6 (24.7;43.4)	0.29 (0.21;0.36)
XGboost	-4.2 (-7.7;-0.5)	-90.6 (-101.3;-82.9)	82.3 (72.3;95.3)	0.67 (0.62;0.71)
biLSTM	-42.2 (-49.3;-35.1)	-162 (-176.9;-149.6)	77.6 (66.3;90.4)	0.4 (0.29;0.49)
10-Min Median - SE (%)				
Decision Tree	-2.1 (-2.8;-1.4)	-18 (-19.9;-16.6)	13.9 (12.6;15.3)	0.22 (0.14;0.29)
Logistic Regression	-21.6 (-22.6;-20.6)	-45.7 (-48.2;-43.8)	2.5 (1;4.3)	0.13 (0.05;0.21)
Feed-Forward Neural Net	-19.1 (-20.1;-18.1)	-42.9 (-44.8;-41.1)	4.7 (3.3;6.2)	0.25 (0.17;0.33)
XGboost	-0.6 (-1.2;-0.1)	-14.5 (-16;-13.3)	13.2 (12.1;14.9)	0.43 (0.36;0.49)
biLSTM	6.4 (5.7;7.2)	-6.5 (-7.8;-5.3)	19.2 (17.7;21.2)	0.16 (0.04;0.27)
10-Min Median - LPS (min)				
Decision Tree	22.8 (17.1;27.6)	-102.7 (-137.1;-83.3)	148.4 (131.9;173.6)	0.06 (-0.02;0.14)
Logistic Regression	60.7 (54.9;65.6)	-64.8 (-100.8;-43.9)	186.2 (168.1;213.6)	0.02 (-0.06;0.1)
Feed-Forward Neural Net	33.8 (28;38.6)	-91.1 (-127.2;-70.2)	158.6 (141.2;184.7)	0.05 (-0.03;0.13)
XGboost	26.4 (21;30.8)	-92.2 (-130;-69.5)	145 (125.5;173.7)	0.1 (0.02;0.18)
biLSTM	-21.5 (-32.7;-12.8)	-187.2 (-253.4;-138.6)	144.3 (104.3;192.6)	0.06 (-0.05;0.18)
10-Min Median - WASO (min)				
Decision Tree	9 (5.2;14.3)	-97.4 (-133.5;-72.1)	115.3 (85.2;163)	0.07 (-0.02;0.15)
Logistic Regression	44.8 (40.8;50)	-66 (-98.3;-45.1)	155.7 (130.2;197.8)	0.17 (0.09;0.25)
Feed-Forward Neural Net	53.4 (49.2;58.7)	-58.6 (-89.6;-38.6)	165.4 (140.4;206.7)	0.22 (0.14;0.3)
XGboost	3.8 (0.3;9.1)	-98.1 (-135.4;-71.9)	105.7 (74.5;153.4)	0.2 (0.13;0.28)
biLSTM	26.8 (19.2;38)	-128.2 (-176.3;-90.8)	181.8 (132.8;250.7)	0.12 (0.01;0.23)

As established from the Table 15, the XGBoost model trained on the 5-minute median filtered ZM data seemed to be the best performing model configuration. The Bland-Altman plot and scatterplot presented in Figure 16 illustrate the level of agreement between the this XGBoost model and ZM-derived sleep quality metrics that were also median-smoothed over 5 minutes. For the sleep quality metrics SPT and TST, the bias is notably close to zero, revealing a minimal average difference with the ZM. The scatterplot for SPT also suggests a moderate linear correlation between the model's predictions and the ZM-derived metrics. The TST scatterplot further indicates a slightly higher correlation, mainly due to the lack of extreme outliers. In contrast, the remaining sleep quality metrics, namely SE, LPS, and WASO, show signs of heteroscedasticity unlike SPT

and TST. While a moderate positive linear correlation exists between the XGBoost model and the ZM-derived SE metrics, poorer correlations are observed for LPS and WASO.

Discussion

In this paper, we assessed a range of machine learning models for estimating in-bed time, sleep time, and derived sleep quality metrics using data from thigh-worn accelerometers. We approached the task both as a hierarchical classification problem, applying models in sequence, and as a multiclass problem using a deep learning model specially tailored for time time series data. These models were trained and assessed using both raw and median-filtered sleep estimates derived from the ZM EEG-based sleep monitor. Overall, all models exhibited strong performance in predicting in bed time on an epoch-to-epoch basis. However, distinguishing between wakefulness and sleep during these in-bed periods proved to be more challenging. Interestingly, while the multiclass biLSTM model performed well in terms of F1 score, precision, and NPV for detecting epoch-to-epoch sleep, it lagged behind in deriving sleep quality metrics when compared to the XGBoost model due to low specificity. The XGBoost model outperformed all included models across every evaluation metric, including epoch-to-epoch prediction and sleep quality metrics. Nonetheless, it's worth noting that all models struggled with relatively low specificity values, indicating a common difficulty in accurately identifying awake epochs during time spent in bed. We also observed performance improvement in all models when 5-minute and 10-minute median filters were applied. This filtering approach resulted in increased total sleep time and sleep efficiency metrics while reducing wake after sleep onset and the number of awakenings. Of all the models, the XGBoost demonstrated the smallest bias and the highest correlation with the ZM sleep quality metrics, making it the most robust choice for this particular application.

While there is limited existing research on the epoch-to-epoch effectiveness of thigh-worn accelerometers in classifying in-bed time, Carlson and colleagues[120] have offered valuable insights. Their study demonstrated that both a third-party algorithm called "ProcessingPal" and a proprietary algorithm, to the activePAL device, named "CREA" were able to achieve high accuracies of 91% and 86%, respectively. Evaluated against self-reported measures in adolescents and adults, these algorithms achieved impressive F1 scores of up to 95% and 96%. These results align with our models, which also achieved over 95% in both F1 and accuracy scores when identifying in-bed epochs, representing time in the SPT. However, it's worth noting that all models in our study, with the exception of XGBoost, tended to underestimate SPT. The biLSTM model displayed the most significant underestimation, with a bias of -36 minutes. This aligns with previous research by Winkler et al.[106], who reported a similar trend in both young-middle-aged and older adults. Their algorithm showed a moderate correlation with diary-recorded waking times but overestimated waking wear time by more than 30 minutes, resulting in an underestimation of in-bed time. This underestimation was further validated by Inan-Eroglu et al.[121], who found an underestimation of 9.8 minutes when comparing Winkler et al.'s algorithm to self-reported measures in middle-aged adults. Contrastingly, another study reported only a slight underestimation of in-bed time in middle-aged and older adults[122]. They used a unique algorithmic approach that quantified the number and

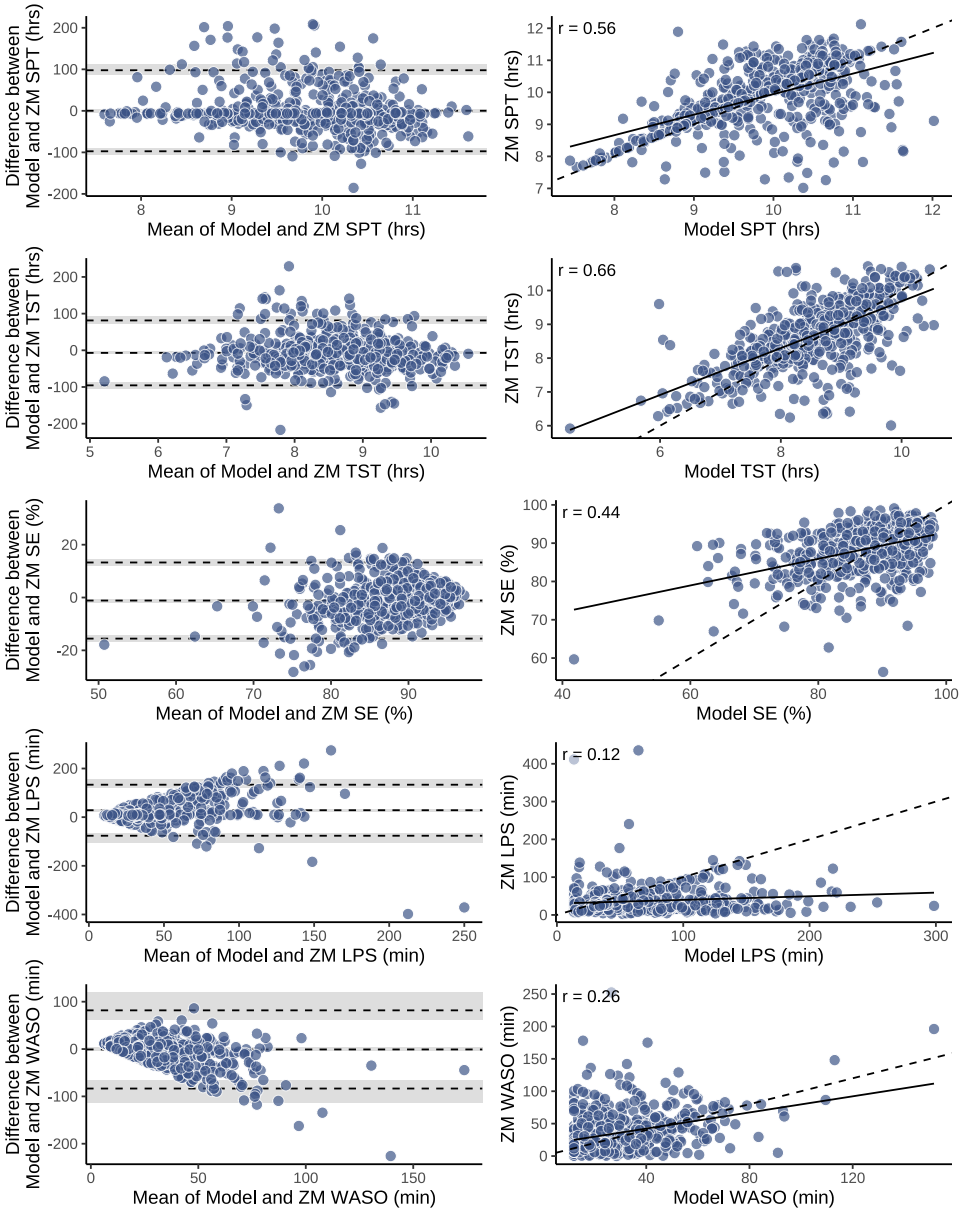

Figure 16: Comparison of sleep quality metrics derived from the XGBoost model trained on the 5-minute smoothed ZM predictions. The left column displays Bland-Altman plots. Dashed lines represent the bias (the average difference between the two measurements) and LOA, with the 95% confidence intervals represented as the grayed areas. The right column displays scatter plots of XGBoost-derived vs ZM-derived sleep quality metrics. The dashed line represents the identity line, while the full-drawn line represents the best linear fit. Pearson's correlations are annotated in the upper left corner

duration of sedentary periods to ascertain time in bed and active periods to identify wake times. Lastly, it's essential to clarify that strong predictive performance in identifying in-bed time doesn't automatically imply accurate predictions for broader sleep quality metrics. Capturing awake periods during in-bed time, a critical factor for assessing other derived sleep quality metrics, isn't effectively handled by simply predicting in-bed time. This distinction between actual sleep and time spent in bed while awake is often overlooked but is vital for a more comprehensive understanding of sleep quality.

To our knowledge, Johansson and colleagues[87] are the only researchers who have gone beyond merely reporting "waking time" and "in-bed time" to provide epoch-to-epoch performance metrics for sleep scoring with thigh-worn accelerometers. Utilizing a single-night evaluation dataset comprising 71 adult subjects, they managed a mean sensitivity of 0.84, a specificity of 0.55, and an accuracy of 0.80. Similarly, our models achieved a high sensitivity of above 97%, but struggled, like Johansson et al.'s algorithm, in detecting in-bed awake epochs. This struggle is manifested in our low specificity scores, which ranged from 54.7% to 76.4%. This challenge is not solely confined to thigh-worn devices. Conley et al.'s[67] meta-analysis reported issues with wrist-worn accelerometers as well, noting mean values of 0.89 for sensitivity, 0.88 for accuracy, and a low 0.53 for specificity among healthy adults. Patterson and colleagues[177] also recently summarized various heuristic algorithms, machine learning, and deep learning models for sleep prediction, finding mean sensitivity and specificity scores of 93% and 60% in data from wrist-worn devices, respectively. These data collectively highlight the persistent difficulty in automating the identification of periods when individuals are awake yet still in bed. Interestingly, our study revealed a divergence from most previous research concerning the overestimation of LPS and WASO by several of our models. This overestimation is reflected in the low NPV scores, indicating that a limited portion of the wake predictions were accurate with the exception of the biLSTM achieving NPV scores up tp 81.8%. This inconsistency might be attributed to the SMOTE we used to balance the dataset. If the synthetic 'wake' samples created by SMOTE don't accurately represent the actual 'wake' data, it could cause the models to misclassify certain 'sleep' epochs as 'wake'. Consequently, this could lead to inflated LPS and WASO estimates, as the models would incorrectly identify more instances of wakefulness during sleep.

The application of the SMOTE in our study likely enhanced the performance of various models by addressing class imbalance issues. However, the introduction of synthetic "wake" samples through this method posed a challenge as they might not be fully indicative of genuine wake data. This could explain why most models, excluding the biLSTM which was not trained on SMOTE-processed data, underestimated TST and SE. Contrarily, the XGBoost model, trained on SMOTE-processed data, managed to navigate these synthetic "wake" samples more effectively and did not overestimate TST as much as other models. This is evident from the Bland-Altman statistics for the XGBoost model trained on 5-minute median-filtered ZM predictions showing a mean difference of -7 minutes for TST and -1.1% for SE. The limits of agreement for these metrics spanned from -95.5 to 81.4 minutes and -15.6% to 13.3% respectively. This suggests that the XGBoost model successfully maintained a balance between sensitivity and specificity without being overly swayed by the synthetic "wake" samples. The robustness of the XGBoost model, especially when faced with synthetic samples, may arise from its gradient boosting mechanism, which involves learning iteratively from the errors of prior models. Unlike other models,

such as the single decision tree, which makes decisions based on impurity measures in one go, or logistic regression, which tries to minimize the loss across the entire dataset, gradient boosting in XGBoost specifically corrects for errors made by previously trained trees with each subsequent tree. Even when compared to iterative algorithms like MLP or biLSTM, XGBoost's approach of directly focusing on and correcting mistakes might make it particularly resilient to inaccuracies or noise introduced by synthetic data. This attribute could lead to better overall model performance.

Sleep detection methods are generally used in two distinct scenarios: night-only recordings and 24-hour recordings. For night recordings, SE and LPS can be readily derived since SPT can be inferred from the length of the recording itself, as indicated by studies from Conley et al. and Patterson et al[67,177]. In contrast, when applied to 24-hour recordings, most sleep detection methods face the challenge of inferring SPT without the aid of sleep diaries, as presented by several studies[42,123,125,178]. This limitation prevents these methods from generating sleep quality metrics dependent on SPT. To address this issue, we designed models capable of distinguishing between in-bed awake time and in-bed asleep time, as well as out-of-bed awake time, over a full 24-hour period. This innovation allows our models to estimate a comprehensive range of commonly used sleep quality metrics. In a similar vein, Van Hees et al.[104] proposed an algorithm for determining SPT using wrist-worn devices, an approach that was subsequently validated by Plekhanova and her team[179]. When combined with other methods, this algorithm enables the estimation of additional sleep quality metrics based on the identified SPT. Van Hees et al. reported favorable results with low mean differences when compared to self-reported measures and PSG for SPT, a finding later corroborated by Plekhanova et al. However, both studies also highlighted challenges in achieving good agreement on metrics such as LPS and WASO, revealing low agreement with PSG. These challenges in accurately detecting wakefulness during in-bed time are similar to the issues we encountered in our own study.

In evaluating various sleep quality metrics, our study identified that LPS consistently exhibited the largest mean error in relation to the actual time allocated to it. This challenge in accurately classifying the initial periods of SPT was further corroborated by poor Pearson correlations between LPS obtained from model predictions and the ZM. Among all models assessed, the XGBoost model emerged as the most reliable, yet it overestimated LPS by an average of 26.4 minutes for models trained on unfiltered ZM predictions. This increased to 28.5 and 34.5 minutes when the training data was 5-minute and 10-minute filtered ZM predictions, respectively. This discrepancy is not unique to our study; it is on par with the mean error of 23 minutes in sleep latency reported by Johansson et al[87]. Johansson et al. suggest that the discrepancy with the gold standard is likely due to the multifaceted nature of the sleep state, which is a complex physiological process. Short awakenings or sleep episodes may not necessarily correspond to noticeable changes in thigh movement, making them difficult to detect and accurately classify. These observations are consistent with findings on wrist-worn devices by Conley and colleagues[67], who reported that correlations between accelerometer data and PSG sleep onset latency (equivalent to LPS) varied greatly across studies. The mean correlation was only 0.2, underscoring the challenges in leveraging accelerometry alone for estimating LPS.

Moreover, when compared to other models like the Van Hees algorithm[82], Oakley rsc[99],

and LSTM-50[99] as evaluated in the Patterson et al. study[177], our XGBoost model displayed narrower LOAs for TST, SE, and WASO. Interestingly, the LOAs were also narrower when pitted against the algorithm tailored for thigh-worn devices by Johansson et al[87], albeit not for SPT. Despite these promising facets, it is important to note that all methods, both from this study and the literature, showcase wide LOAs. This implies a high level of variability in sleep quality metrics derived from accelerometry, cautioning against its use as a stand-alone alternative to EEG-based ZM or PSG for individual-level sleep assessments. The presence of extreme outliers in our study appeared to exacerbate the width of the LOAs, suggesting that current methods are better suited for group-level sleep quality metrics. As a result, there is a pressing need for further refinement to enhance the reliability and validity of these models for individual sleep assessments.

In our study, we opted to use the ZM as the reference method for sleep measurement, as opposed to the generally accepted gold standard, PSG. While this choice could contribute to the observed discrepancies between our models and ZM outcomes, we argue that the use of ZM has distinct advantages. For one, ZM facilitates multiple consecutive nights of recording in a free-living environment[52], thereby capturing intra-individual variations in sleep patterns. This is an aspect often impractical to achieve with PSG. Additionally, the use of ZM allowed us to incorporate more nights into our study than is typically possible with PSG-based studies. This is evident when comparing our data set to the more limited Newcastle dataset, which consists of only 28 participants each with a single night of recoring[180]. Despite its benefits, we found that the raw ZM outputs were not ideally suited for developing machine learning models, primarily due to a low signal-to-noise ratio, as indicated in Figure 13. The ZM device itself employs certain filtering processes to mitigate this issue when generating sleep quality metrics. For example, WASO is determined using contiguous epochs of 3 minutes, and sleep only contributes to sleep quality metrics if 10 out of 12 minutes are categorized as sleep. To enhance the effectiveness of our machine learning algorithms, we applied median filters to the raw ZM predictions, which had a notable impact on the derived sleep quality metrics. The application of these filters led to several changes. Specifically, the mean WASO dropped from 39 minutes in the raw predictions to 30.6 minutes when using a 5-minute median filter, and further decreased to 22.3 minutes with a 10-minute median filter. Likewise, TST, SE, and LPS all increased upon the application of the 5-minute and 10-minute median filters. These shifts suggest that the median filters reclassify brief instances of wakefulness as sleep, and similarly, eliminate short awakenings. Despite these alterations, the sleep quality metrics derived from the median-filtered predictions remained largely consistent with those from the raw predictions, validating the approach we took in this study.

Our current study offers contributions to the field of sleep estimation methods, particularly in the use of thigh-worn accelerometers. One of the primary strengths of the research lies in its ability to distinguish between in-bed awake time and in-bed asleep time, as well as out-of-bed time. This distinction is crucial for extracting essential sleep quality metrics without the assistance of sleep diaries or other ways to determine the SPT. Additionally, by evaluating multiple nights per subject, the study offers valuable insights into intra-subject sleep variability, another important factor for sleep assessment. However, there are limitations to our approach. Most notably, we utilized the ZM as our reference method, which is not considered the gold standard for sleep measurement.

This choice might have impacted the validity of our findings. For future work, it would be beneficial to employ PSG as a reference, despite its own set of limitations, to provide a more accurate comparison and to easier extend to comparisons of other studies utilizing PSG as the gold standard. Another limitation is the lack of external validation for our models, which confines the applicability of our findings primarily to child populations of normal sleepers. Finally, another significant limitations of importantce is that the sequential strategy, during its initial step, identifies only one continuous SPT window. Consequently, it cannot detect multiple SPTs in a single night. So, if participants wake up for an extended period, engage in significant physical activity, and then return to sleep later, the detected SPT will terminate at the wake point, and any subsequent sleep will not be recognized.

Thesis Conclusions

The importance of sleep health and the limitations of conventional PSG examinations have highlighted the potential of wearable sensor systems as complementary measurement methods. In the previous sections of this thesis, it has been shown that accelerometers as a platform for long-term activity monitoring can be used to assess quantitative and qualitative aspects of sleep. Specifically, the findings of this thesis offer a deep dive into the methods and tools employed in the manual annotation of in-bed periods, the classification of non-wear periods, and the application of various machine learning models to detect sleep in accelerometry data.

Regarding the objectives set in the beginning of the thesis, it can be stated that:

- We presented a method using Audacity for manually annotating individual bedtime and wake-up times in raw accelerometry data from thigh- and hip-worn accelerometers. The annotations showed good to excellent absolute agreement with in-bed and out-of-bed timestamps as determined by the ZM sleep tracking device and with prospective sleep diaries, addressing the objective of validating the accuracy of these annotations. Additionally, the manual annotations displayed excellent inter- and intra-rater agreement between the three human raters and between test/retest conditions.
- In alignment with the first objective of Paper II, we evaluated decision tree models using data from thigh and hip-worn accelerometers to detect non-wear time, with a focus on the role of surface skin temperature. Our findings highlighted the strong performance of these models, particularly when incorporating surface skin temperature, closely mirroring the effectiveness of the simple hueristic algorithms, `cz_60` and `heu_alg`, for non-wear episodes longer than 60 minutes across all wear-locations: wrist, hip, and thigh. Addressing the second objective, our comparative analysis showed notable performance distinctions among machine-learned models and heuristic algorithms. Specifically, the decision tree model trained on the six most important predictors excelled for short non-wear episodes on all wear-locations, while some the more complex models, the random forest (`sunda_RF`) and deep learning model (`syed_CNN`), faced limitations, especially in identifying shorter non-wear episodes.
- In accordance with the primary objective of Paper III, we evaluated various machine learning and deep learning models to estimate in-bed and sleep time, benchmarked against sleep recordings of the ZM. All models showcased strong performance in predicting in-bed time on an epoch-to-epoch basis, albeit with challenges in distinguishing wakefulness and sleep during these periods. The secondary objective entailed assessing the models' capability in quantifying commonly used sleep quality metrics, validated against the ZM sleep tracking device. Among the evaluated models, the XGBoost model excelled across every evaluation metric, including epoch-to-epoch prediction and sleep quality metrics, outperforming others like the multiclass biLSTM model, especially in deriving sleep quality metrics due to its higher specificity. The application of 5-minute and 10-minute median filters

resulted in a performance uplift for all models, although these filterings at the same time increased predicted time spent asleep, thus, increasing sleep quality metrics such as TST and SE while reducing WASO and the number of awakenings. The XGBoost model, exhibiting the smallest bias and highest correlation with the ZM sleep quality metrics, emerged as the most robust choice for this application.

Perspectives

While the findings of this thesis has bridged knowledge gaps, further horizons beckon exploration. The following will discuss implications of the findings in this thesis and potential avenues for future reseach.

Implications of Findings

As demonstrated in Paper I, the manual annotation method described for labeling in-bed periods using raw accelerometry data eliminates the need for sophisticated preprocessing to extract meaningful sleep information. This manual annotation serves as a valid and consistent alternative to sleep diaries or in-bed time determined by EEG recordings. Accurate and correct labeling is paramount, not just in sleep research, but across many fields, as it forms the foundation for effective machine learning applications. Reliable labels, such as those provided by manual annotations, allow for more accurate model training, which in turn leads to better predictions and insights. An accelerometry dataset enhanced with these manual annotations—especially in the absence of accompanying sleep diaries or other in-bed time estimates—can streamline and expedite data analysis workflows in sleep research, resulting in increased efficiency.

Through the utilization of simple decision tree models and by recognizing the role of surface skin temperature, as presented in Paper II, we have advanced in our capability to accurately identify non-wear time. This may reduce potential data inaccuracies and biases towards excessive sleep or sedentary time, ensuring the integrity of our accelerometry datasets. Additionally, our detailed evaluation of different wear locations—including the hip, thigh, and wrist—offers researchers critical insights into the efficacy variances of different methods based on the device's position. We've also placed significant emphasis on the value of external validation, highlighting the essential need for stringent validation protocols before the declaration of model performance. Our findings clearly indicate that heuristic algorithms stand out as both population and wear-time agnostic. They offer enhanced flexibility in study design. However, when it comes to detecting shorter non-wear episodes—a domain where traditional algorithms often falter—it's worth considering machine learning model-based approaches but such should be carefully selected.

By assessing a range of machine learning and deep learning models against an EEG-based sleep monitor in Paper III, we've gained critical insights into the algorithms most apt for precise sleep detection given the employed feature set and data at hand. This knowledge promotes the selection of the most effective models, enhancing the precision of sleep detection from accelerometry-based devices. Moreover, these validated methods using accelerometers not only facilitate sleep detection without the need for sleep logs or diaries but also allow for easy assessment of previously collected data for sleep patterns. These approaches offer a promising, cost-efficient, and less obtrusive replacement for traditional EEG-based sleep monitoring methods. By doing so, they can make sleep

studies more universally accessible, ensuring that a broader spectrum of the population can benefit from and participate in such studies.

Conclusively, the research presented in this thesis, while significant in its attributions, reinforces a timeless academic tradition: with each answer, new questions arise, directing future research in the intersection of sleep health and technology.

Initial Study Design Considerations

Research is inherently iterative, building upon existing knowledge while continuously adapting based on insights and practical experiences. At the outset of our project, we intended to harness the advanced capabilities of the SomnoTouch™ RESP[181], an FDA-approved ambulatory PSG system, to monitor in-home sleep patterns alongside concurrent accelerometry assessments. Our goal was to create a comprehensive, multi-dimensional dataset. With manual annotation of in-bed periods, as outlined in Paper I, we aimed to establish a gold standard dataset, laying the groundwork for training various machine learning models dedicated to sleep detection. Yet, as is customary in research, we sometimes encounter roadblocks when translating theoretical designs into practical applications. Using the SomnoTouch™ RESP for children's in-home monitoring proved challenging. Despite the device's advanced features and accuracy, achieving consistent and comfortable adherence of its many sensors with children presented substantial hurdles. Often, we were only able to obtain a few hours of high-quality data before losing connectivity with the electrodes, typically because the children were lying restlessly in bed while awake or experiencing restless sleep later on. Additionally, there were several instances where the signal was lost as the children went to bed, likely due to the 'diving head first into the pillow' effect.

Before terminating my data collection, I collected data from 55 children, employing stronger EEG-electrode adherence paste with each trial and providing progressively stricter instructions on how to carefully manage the wiring and electrodes when going to bed. Despite these measures, the data quality remained subpar. Such challenges prompted a reevaluation and adaptation of our initial methodology. These experiences underscore the importance of research tools being versatile, not just in their precision, but also in accommodating the peculiarities of the target population and environment. Future trials with improved attachment methods, or perhaps utilizing an EEG helmet, could be beneficial. However, a balance needs to be struck: how many sensors and how much attachment equipment can a child comfortably sleep with before it ceases to represent natural, free-living sleep? Ultimately, the ideal dataset for sleep research would seamlessly blend free-living accelerometer data with gold standard sleep recordings over multiple days. Although attaining this high-quality dataset is challenging, it's pivotal for advancing sleep research.

Addressing Non-Wear Challenges

In physical activity research, accurate monitoring with wearable devices is essential, especially for non-wear detection. Misclassifying non-wear periods can lead to overestimated assessments, altering our understanding of actual sleep or sedentary behaviors. Such inaccuracies might affect the trust in wearable devices for clinical purposes. This accuracy becomes particularly vital in long-term studies spanning several weeks. Participants may intermittently remove their devices during these periods, leading to varied non-wear breaks, making accurate detection crucial.

In Paper II, while our data showed a dominance of non-wear episodes longer than 60 minutes, other research points in the opposite direction[113,129,149,153]. This variation stresses the need for refining existing non-wear detection algorithms. For instance, Zhou et al.'s algorithm[117] was based on 15-minute minimum reference periods for wear and non-wear. Although the researchers acknowledged limitations for shorter durations, they believed the broader implications were minimal. Yet, errors in identifying brief non-wear periods could skew individual outcomes, especially in longitudinal studies where these errors accumulate.

With the shift to raw accelerometer data for non-wear detection, we face challenges and uncertainties in data management and processing. One significant concern is the sampling frequency. Algorithms are often designed for specific sampling rates, and using them on data with different frequencies could lead to errors. The nuances of resampling to different frequencies remain murky. Interpolative sampling algorithms, with their inherent assumptions, might not always provide accurate results. It's essential to understand how this affects accelerometer values. Additionally, accelerometer calibration, typically done during manufacturing, plays a role. Sometimes, recalibrating data, as with methods like auto-calibration[182], can be beneficial, and understanding its impact on non-wear detection is crucial.

The ongoing challenge is crafting algorithms that can consistently detect non-wear across different devices, user demographics, and wear locations. If a one-size-fits-all model proves elusive, specialized models for specific scenarios might be the solution. Paper II underlined the difficulties in creating universal methods for short non-wear periods. However, by incorporating skin temperature data, we achieved commendable F1 scores across wear locations, even though some of the results may not meet the threshold for individual reliability.

Improving Our Sleep Models

In Paper III, we highlighted a limitation associated with our model's ability to detect only one SPT per night. Given the specific sample we analyzed, the detected SPTs ranged between 7.4 to 12 hours, with an average of 9.9 hours and a standard deviation of 0.8 hours, as determined by the XGBoost model. Given this range, it's improbable that our dataset experienced interrupted SPTs. However, it's worth noting that the nights under study were meticulously selected based on the ZM recordings to exclude instances with abnormally short or extended SPTs. In different datasets, there's potential for some nights

to be truncated due to extended periods of wakefulness characterized by significant physical activity, or even napping behaviors. This limitation is a noteworthy drawback in our methodology, warranting attention in future revisions. Addressing this concern and allowing for the detection of multiple SPTs should be a feasible programming enhancement, though it was not the primary focus during the development of our methodology. Daytime sleep behavior is another area of potential research interest. While the health implications of nocturnal insufficient and poor sleep are well-established, there's evidence indicating an increased mortality associated with habitual daytime sleep[183]. However, other findings have shown potential benefits of daytime napping, particularly in enhancing glycaemic control for individuals with type 2 diabetes[184]. Daytime sleep could also influence the relationship between physical activity and sleep. Consequently, forthcoming research could emphasize the creation of algorithms adept at detecting multiple sleep periods, capturing both daytime naps and nocturnal sleep.

As observed in Table 12, the variation in SPTs as identified by the ZM (mean = 9.2 hours, sd = 2.1) exceeds that of the SPTs determined by the XGBoost model. This disparity may stem from the disproportionate emphasis the model places on circadian rhythm features (clock proxy features), which might inadvertently limit its ability to recognize more diverse SPTs. Future investigations should delve into this matter, perhaps refining the features related to circadian rhythms to ensure a more balanced and accurate model output. Moreover, expanding the feature set could offer a potential solution, allowing the model to capture a wider range of sleep patterns and behaviors. Additionally, it would be prudent to explore more comprehensive biLSTM architectures. Integrating the biLSTM model in a sequential manner, similar to the approach taken with the XGBoost models, might also yield improved results and merits further investigation. Finally, the biLSTM would likely benefit from another loss function like a soft version of the F1 score like described by Pastor-Pellicer et al.[185] to better account for the class imbalance.

As presented in this thesis, the estimation of sleep, as evidenced by the limits of agreements, is not sufficiently reliable to be conducted on an individual basis. Looking beyond the present scope of our research, there lies an intriguing possibility in the development of adaptive machine learning models that tailor predictions based on individual sleep patterns[186]. Such models, known for their real-time adaptability, could cater to outliers or individuals with specific sleep disorders, ensuring personalized insights. This adaptive approach presents an exciting avenue for future sleep research, and while it may seem distant, the potential benefits in terms of accuracy and personalization are significant.

Generalizability of Our Sleep Models

In Paper III, we focused on a sample population of children with a mean age of 9.2 years and a standard deviation of 2.1. By focusing on a pediatric demographic, our study provides unique insights into the sleep patterns and habits of children, a population that may sometimes be underrepresented in sleep studies. As children undergo significant developmental changes, understanding their sleep behavior is crucial. However, to ensure our algorithm's utility across the life course, the need for diverse validation becomes all the more imperative. As noted by Mukherjee et al.[187], sleep physiology and

habits evolve throughout an individual's life. This brings forward a critical consideration about the broader applicability of our algorithm across different age groups and clinical demographics. While studies have indicated that algorithms tailored for adults might not always align with children's sleep patterns, often detecting wakefulness when they are likely asleep[188], our study was already grounded in pediatric sleep patterns. Nevertheless, the elderly and certain clinical populations exhibit distinct sleep profiles, often characterized by sleep-related challenges[189,190]. Conditions such as obesity and type 2 diabetes, which are associated with a heightened risk of obstructive sleep apnoea[191,192], could introduce variations in sleep detection. The algorithm developed by van Hees et al. highlighted performance disparities between healthy sleepers and those with sleep disorders[104]. Given our findings within the pediatric demographic, future work should involve the validation of our algorithm across a broader age spectrum and in varied clinical contexts to ensure its comprehensive reliability.

Transitioning to the topic of external validation, all splits of data for Paper III was sourced from one specific research project. To arrive at more concrete conclusions and to fortify the model's generalizability, diversifying datasets for external validation in future studies is important. However, the inherent challenge remains that amassing such data is a difficult task, demanding significant time and resources. Especially as there, to my knowledge, does not exist any publicly available datasets from thigh-worn accelerometers coupled with valid sleep records like the Newcastle or MESA datasets[180,193,194]. It is thus essential to recognize the potential need for data that encapsulates diverse sleep behaviors, including napping, to ensure comprehensive sleep detection in future research.

Multimodal Sensor Integration

Every year sees the release of an increasing number of devices. While the accuracy of early commercial devices faced skepticism based on scientific evaluations, recent multisensor devices now showcase precision that parallels leading research tools[195,196]. A prime example of advancements in consumer sleep wearables is the Oura-Ring. Designed to be worn on the finger, this device combines the functionalities of an accelerometer and a pulse-oximeter. The first-generation Oura Ring detected sleep with a sensitivity of 96% and was able to pinpoint REM sleep with an accuracy of 61% among adolescents and young adults[195]. Its successor, the second generation, touted an accuracy rate of 94% in sleep detection using accelerometer-based features and reached 96% accuracy when augmented with autonomic nervous system-derived features[197]. Using a four-stage model for sleep detection—which encompasses light sleep (N1 and N2 stages), deep sleep (N3), REM, and wake—the device managed an accuracy of 57% with accelerometer-derived data and jumped to 79% when oximeter data was included. However, performance varies across devices. For instance, various Fitbit Wristband models exhibit differing accuracies for specific sleep stages: they range between 69%-81% for light sleep, 36%-89% for deep sleep, and 62%-89% for REM sleep[@haghayegh_2019]. Such fluctuations indicate that relying solely on acceleration data may not always yield a comprehensive view of sleep stages. Further evidence points to heart rate variability (HRV) as a significant marker for sleep stage detection[198,199]. A notable study achieved a striking 89% accuracy in identifying deep sleep when HRV was combined with respiratory signals[200]. Another

research effort merged HRV with accelerometer data, achieving accurate detection for 75% of deep sleep and over 70% of REM sleep instances, though light sleep identification varied between 42% and 52%[201]. This thesis, however, zeroed in on algorithms rooted in acceleration and surface skin temperature data. While the inclusion of heart rate and other physiological markers could enhance sleep classification, this enhancement isn't without its trade-offs.

Choosing between a streamlined sensor system and a robust multi-sensor setup largely depends on the specific application. Short-term sleep assessments, typically spanning a week or two for research or clinical purposes, may benefit from a broader sensor array, given the depth of information it offers. Conversely, for longer evaluations stretching over months, prioritizing comfort and minimizing data loss risks, whether due to technical hitches or user compliance, gains prominence. Even with the integration of multimodal sensors, this approach remains cost-efficient compared to PSG and also opens doors to expansive research opportunities and interventions in sleep science.

As a closing remark, the journey embarked upon in this thesis is not a culmination but a commencement. The vision is clear: to reenvision sleep health, making it more accessible, accurate, and actionable for society at large. As technology and healthcare coalesce, the dream of a well-rested world, empowered by data-driven insights, inches ever closer to reality.

Code Availability

All code associated with Paper II, used for data processing and analysis, and for producing figures, tables, and results, is available at https://github.com/esbenlykke/nonwear_project. This specific codebase requires substantial refactoring and commenting to adhere to best coding practices. However, upon request, the corresponding author can assist in utilizing the code if needed.

For Paper III, all code used for data processing, analysis, figure and table generation, results production, and manuscript pdf document creation is available at https://github.com/esbenlykke/sleep_study. This codebase largely adheres to good coding practices and should be deployable 'out of the box', provided all necessary dependencies are installed. To facilitate this, I plan to provide a conda/mamba environment description in the repository to encompass all required dependencies. Note that the repository currently contains some redundant scripts; however, I intend to streamline the content in the future. Additionally, plans are underway to introduce a Snakemake file to automate the entire process.

I've also developed a tool based on the findings from Paper III. This tool leverages the best-performing models (specifically, the XGBoost models trained on 5-minute median-filtered ZM sleep predictions) and is available at https://github.com/esbenlykke/get_sleep_stats. When given the path to a folder with .wav or .cwa raw accelerometer files, the tool first extracts all relevant features from the raw data. It then predicts and extracts in-bed periods. Lastly, it predicts sleep time based on the extracted in-bed periods. The final output includes timestamps for going to bed and leaving bed, SPT, TST, and SE for each accelerometer recording.

No code is available for Paper I.

References

1. Kraus, W. E. et al. Physical activity, all-cause and cardiovascular mortality, and cardiovascular disease. Med Sci Sports Exerc **51**, 1270–1281 (2019).
2. Lee, I.-M. et al. Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. Lancet **380**, 219–229 (2012).
3. Wilmot, E. G. et al. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: Systematic review and meta-analysis. Diabetologia **55**, 2895–2905 (2012).
4. Cappuccio, F. P., D'Elia, L., Strazzullo, P. & Miller, M. A. Sleep duration and all-cause mortality: A systematic review and meta-analysis of prospective studies. Sleep **33**, 585–592 (2010).
5. Jennum, P. et al. Søvn og sundhed | Vidensråd for Forebyggelse.
6. Piercy, K. L. et al. The physical activity guidelines for americans. JAMA **320**, (2018).
7. Ahrensberg, H., Toftager, M., Nørgaard, S., Petersen, C. B. & Larsen, C. V. Fysisk aktivitet for voksne (18-64 år): Viden om forebyggelse og sundhed. (Sundhedsstyrelsen, 2023).
8. Ahrensberg, H., Toftager, M., Petersen, C. B. & Wehner, S. K. Fysisk aktivitet for børn og unge (5-17 år): Viden om sundhed og forebyggelse. (Sundhedsstyrelsen, 2023).
9. Hirshkowitz, M. et al. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. Sleep Health **1**, 40–43 (2015).
10. Paruthi, S. et al. Consensus statement of the american academy of sleep medicine on the recommended amount of sleep for healthy children: Methodology and discussion. J Clin Sleep Med **12**, 1549–1561 (2016).
11. Watson, N. F. et al. Recommended amount of sleep for a healthy adult: A joint consensus statement of the american academy of sleep medicine and sleep research society. Sleep **38**, 843–844 (2015).
12. Ma, G. Sleep, health, and society. Sleep medicine clinics **12**, (2017).
13. Worley, S. L. The Extraordinary Importance of Sleep: The Detrimental Effects of Inadequate Sleep on Health and Public Safety Drive an Explosion of Sleep Research. P & T: A Peer-Reviewed Journal for Formulary Management **43**, 758–763 (2018).
14. Matricciani, L., Paquet, C., Galland, B., Short, M. & Olds, T. Children's sleep and health: A meta-review. Sleep Medicine Reviews **46**, 136–150 (2019).
15. Scott, A. J., Webb, T. L., Martyn-St James, M., Rowse, G. & Weich, S. Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. Sleep Medicine Reviews **60**, 101556 (2021).

16. Consensus Conference Panel et al. Recommended amount of sleep for a healthy adult: A joint consensus statement of the american academy of sleep medicine and sleep research society. J Clin Sleep Med **11**, 591–592 (2015).

17. Ji, A. et al. Interactive effect of sleep duration and sleep quality on risk of stroke: An 8-year follow-up study in China. Scientific Reports **10**, 8690 (2020).

18. Hale, L., Troxel, W. & Buysse, D. J. Sleep Health: An Opportunity for Public Health to Address Health Equity. Annual Review of Public Health **41**, 81–99 (2020).

19. Shochat, T., Cohen-Zion, M. & Tzischinsky, O. Functional consequences of inadequate sleep in adolescents: a systematic review. Sleep Medicine Reviews **18**, 75–87 (2014).

20. Kecklund, G. & Axelsson, J. Health consequences of shift work and insufficient sleep. BMJ (Clinical research ed.) **355**, i5210 (2016).

21. O'Brien, E. M. & Mindell, J. A. Sleep and risk-taking behavior in adolescents. Behavioral Sleep Medicine **3**, 113–133 (2005).

22. Bonnet, M. H. Effect of sleep disruption on sleep, performance, and mood. Sleep **8**, 11–19 (1985).

23. Connor, J. et al. Driver sleepiness and risk of serious injury to car occupants: population based case control study. BMJ (Clinical research ed.) **324**, 1125 (2002).

24. Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A. & Bögels, S. M. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. Sleep Medicine Reviews **14**, 179–189 (2010).

25. Roth, T. & Roehrs, T. A. Etiologies and sequelae of excessive daytime sleepiness. Clinical Therapeutics **18**, 562–576 (1996).

26. Wang, H. et al. Genome-wide association analysis of self-reported daytime sleepiness identifies 42 loci that suggest biological subtypes. Nature Communications **10**, 3503 (2019).

27. Rollo, S., Antsygina, O. & Tremblay, M. S. The whole day matters: Understanding 24-hour movement guideline adherence and relationships with health indicators across the lifespan. Journal of Sport and Health Science **9**, 493–510 (2020).

28. Roebuck, A. et al. A review of signals used in sleep analysis. Physiological measurement **35**, R1–57 (2014).

29. Stamatakis, K. A. & Punjabi, N. M. Effects of sleep fragmentation on glucose metabolism in normal subjects. Chest **137**, 95–101 (2010).

30. Herzog, N. et al. Selective slow wave sleep but not rapid eye movement sleep suppression impairs morning glucose tolerance in healthy men. Psychoneuroendocrinology **38**, 2075–2082 (2013).

31. Reutrakul, S. & Van Cauter, E. Sleep influences on obesity, insulin resistance, and risk of type 2 diabetes. Metabolism: Clinical and Experimental **84**, 56–66 (2018).

32. Cappuccio, F. P., Cooper, D., D'Elia, L., Strazzullo, P. & Miller, M. A. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. European Heart Journal **32**, 1484–1492 (2011).

33. João, K. A. D. R., Jesus, S. N. de, Carmo, C. & Pinto, P. The impact of sleep quality on the mental health of a non-clinical population. Sleep Medicine **46**, 69–73 (2018).

34. Cappuccio, F. P. et al. Meta-analysis of short sleep duration and obesity in children and adults. Sleep **31**, 619–626 (2008).

35. Banks, S. & Dinges, D. F. Behavioral and physiological consequences of sleep restriction. Journal of Clinical Sleep Medicine : JCSM : official publication of the American Academy of Sleep Medicine **3**, 519–528 (2007).

36. Van Cauter, E., Spiegel, K., Tasali, E. & Leproult, R. Metabolic consequences of sleep and sleep loss. Sleep medicine **9**, S23–S28 (2008).

37. Buysse, D. J. Sleep health: Can we define it? Does it matter? Sleep **37**, 9–17 (2014).

38. Basnet, S. et al. Associations of common chronic non-communicable diseases and medical conditions with sleep-related problems in a population-based health examination study. Sleep Science **9**, 249–254 (2016).

39. Aserinsky, E. & Kleitman, N. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. Science (New York, N.Y.) **118**, 273–274 (1953).

40. Sadeh, A. Iii. Sleep Assessment Methods. Monographs of the Society for Research in Child Development **80**, 33–48 (2015).

41. Ibáñez, V., Silva, J. & Cauli, O. A survey on sleep assessment methods. PeerJ **6**, e4849 (2018).

42. Girschik, J., Fritschi, L., Heyworth, J. & Waters, F. Validation of self-reported sleep against actigraphy. J Epidemiol **22**, 462–468 (2012).

43. Levendowski, D. J. et al. The Accuracy, Night-to-Night Variability, and Stability of Frontopolar Sleep Electroencephalography Biomarkers. Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine **13**, 791–803 (2017).

44. Redline, S., Dean, D. & Sanders, M. H. Entering the era of "big data": Getting our metrics right. Sleep **36**, 465–469 (2013).

45. Berthomier, C. & Brandewinder, M. Sleep scoring: man vs. machine? Sleep and Breathing **17**, 461–462 (2013).

46. Malhotra, A. et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep **36**, (2013).

47. Koley, B. & Dey, D. An ensemble system for automatic sleep stage classification using single channel EEG signal. Computers in Biology and Medicine **42**, 1186–1195 (2012).

48. Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H. & Dickhaus, H. Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. Computer Methods and Programs in Biomedicine **108**, 10–19 (2012).

49. Zhu, G., Li, Y. & Wen, P. P. Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. IEEE journal of biomedical and health informatics **18**, 1813–1821 (2014).

50. Kaplan, R. F., Wang, Y., Loparo, K. A., Kelly, M. R. & Bootzin, R. R. Performance evaluation of an automated single-channel sleep–wake detection algorithm. Nat Sci Sleep **6**, 113–122 (2014).

51. Wang, Y., Loparo, K. A., Kelly, M. R. & Kaplan, R. F. Evaluation of an automated single-channel sleep staging algorithm. Nat Sci Sleep **7**, 101–111 (2015).

52. Pedersen, J., Rasmussen, M. G. B., Olesen, L. G., Kristensen, P. L. & Grøntved, A. Self-administered electroencephalography-based sleep assessment: Compliance and perceived feasibility in children and adults. Sleep Science and Practice **5**, 8 (2021).

53. Silva, G. E., Vana, K. D., Goodwin, J. L., Sherrill, D. L. & Quan, S. F. Identification of patients with sleep disordered breathing: Comparing the four-variable screening tool, STOP, STOP-bang, and epworth sleepiness scales. Journal of Clinical Sleep Medicine **07**, 467–472 (2011).

54. El-Sayed, I. H. Comparison of four sleep questionnaires for screening obstructive sleep apnea. Egyptian Journal of Chest Diseases and Tuberculosis **61**, 433–441 (2012).

55. Firat, H., Yuceege, M., Demir, A. & Ardic, S. Comparison of four established questionnaires to identify highway bus drivers at risk for obstructive sleep apnea in Turkey. Sleep and Biological Rhythms **10**, 231–236 (2012).

56. Luo, J., Huang, R., Zhong, X., Xiao, Y. & Zhou, J. STOP-Bang questionnaire is superior to Epworth sleepiness scales, Berlin questionnaire, and STOP questionnaire in screening obstructive sleep apnea hypopnea syndrome patients. Chinese Medical Journal **127**, 3065–3070 (2014).

57. Pataka, A., Daskalopoulou, E., Kalamaras, G., Fekete Passa, K. & Argyropoulou, P. Evaluation of five different questionnaires for assessing sleep apnea syndrome in a sleep clinic. Sleep Medicine **15**, 776–781 (2014).

58. Chasens, E. R., Ratcliffe, S. J. & Weaver, T. E. Development of the FOSQ-10: A short version of the functional outcomes of sleep questionnaire. Sleep **32**, 915–919 (2009).

59. Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K. & Rathouz, P. J. Sleep duration: How well do self-reports reflect objective measures? The CARDIA sleep study. Epidemiology (Cambridge, Mass.) **19**, 838–845 (2008).

60. Thurman, S. M. et al. Individual differences in compliance and agreement for sleep logs and wrist actigraphy: A longitudinal study of naturalistic sleep in healthy adults. PloS One **13**, e0191883 (2018).

61. Migueles, J. H. et al. Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. Sports Med **47**, 1821–1845 (2017).

62. Heesch, K. C., Hill, R. L., Aguilar-Farias, N., Uffelen, J. G. Z. van & Pavey, T. Validity of objective methods for measuring sedentary behaviour in older adults: a systematic review. The International Journal of Behavioral Nutrition and Physical Activity **15**, 119 (2018).

63. Skotte, J., Korshøj, M., Kristiansen, J., Hanisch, C. & Holtermann, A. Detection of physical activity types using triaxial accelerometers. Journal of Physical Activity and Health **11**, 76–84 (2014).

64. Brønd, J. C., Grøntved, A., Andersen, L. B., Arvidsson, D. & Olesen, L. G. Simple Method for the Objective Activity Type Assessment with Preschoolers, Children and Adolescents. Children (Basel, Switzerland) **7**, 72 (2020).

65. Arvidsson, D. et al. Re-examination of accelerometer data processing and calibration for the assessment of physical activity intensity. Scand J Med Sci Sports **29**, 1442–1452 (2019).

66. Chen, K. Y. & Bassett, D. R. The technology of accelerometry-based activity monitors: current and future. Medicine and Science in Sports and Exercise **37**, S490–500 (2005).

67. Conley, S. et al. Agreement between actigraphic and polysomnographic measures of sleep in adults with and without chronic conditions: A systematic review and meta-analysis. Sleep Medicine Reviews **46**, 151–160 (2019).

68. Meredith-Jones, K., Williams, S., Galland, B., Kennedy, G. & Taylor, R. 24 h Accelerometry: impact of sleep-screening methods on estimates of sedentary behaviour and physical activity while awake. Journal of Sports Sciences **34**, 679–685 (2016).

69. Tudor-Locke, C., Camhi, S. M. & Troiano, R. P. A catalog of rules, variables, and definitions applied to accelerometer data in the National Health and Nutrition Examination Survey, 2003-2006. Preventing Chronic Disease **9**, E113 (2012).

70. Sadeh, A., Sharkey, K. M. & Carskadon, M. A. Activity-based sleep-wake identification: An empirical test of methodological issues. Sleep **17**, 201–207 (1994).

71. Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J. & Gillin, J. C. Automatic sleep/wake identification from wrist activity. Sleep **15**, 461–469 (1992).

72. Hjorth, M. F. et al. Measure of sleep and physical activity by a single accelerometer: Can a waist-worn actigraph adequately measure sleep in children? Sleep and Biological Rhythms **10**, 328–335 (2012).

73. Tilmanne, J., Urbain, J., Kothare, M. V., Wouwer, A. V. & Kothare, S. V. Algorithms for sleep-wake identification using actigraphy: a comparative study and new results. Journal of Sleep Research **18**, 85–98 (2009).

74. Souza, L. de et al. Further validation of actigraphy for sleep studies. Sleep **26**, 81–85 (2003).

75. Littner, M. et al. Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002. Sleep **26**, 337–341 (2003).

76. Sazonov, E., Sazonova, N., Schuckers, S., Neuman, M. & CHIME Study Group. Activity-based sleep-wake identification in infants. Physiol Meas **25**, 1291–1304 (2004).

77. Granovsky, L., Shalev, G., Yacovzada, N.-S., Frank, Y. & Fine, S. Actigraphy-based sleep/wake pattern detection using convolutional neural networks. ArXiv (2018).

78. Smith, M. T. et al. Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Clinical Practice Guideline. Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine **14**, 1231–1237 (2018).

79. Neishabouri, A. et al. Quantification of acceleration as activity counts in ActiGraph wearable. Scientific Reports **12**, 11958 (2022).

80. Webster, J. B., Kripke, D. F., Messin, S., Mullaney, D. J. & Wyborney, G. An activity-based sleep monitor system for ambulatory use. Sleep **5**, 389–399 (1982).

81. Borazio, M., Berlin, E., Kucukyildiz, N., Scholl, P. & Van Laerhoven, K. 2014 IEEE international conference on healthcare informatics. in 125–134 (2014).

82. Hees, V. T. van et al. A novel, open access method to assess sleep duration using a wrist-worn accelerometer. PLOS ONE **10**, e0142533 (2015).

83. Difrancesco, S. et al. Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study. Depression and Anxiety **36**, 975–986 (2019).

84. Jones, S. E. et al. Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. Nature Communications **10**, 1585 (2019).

85. Koopman-Verhoeff, M. E. et al. Preschool family irregularity and the development of sleep problems in childhood: a longitudinal study. Journal of Child Psychology and Psychiatry **60**, 857–865 (2019).

86. Häusler, N., Marques-Vidal, P., Haba-Rubio, J. & Heinzer, R. Association between actigraphy-based sleep duration variability and cardiovascular risk factors – results of a population-based study. Sleep Medicine **66**, 286–290 (2020).

87. Johansson, P. J. et al. Development and performance of a sleep estimation algorithm using a single accelerometer placed on the thigh: An evaluation against polysomnography. Journal of Sleep Research **32**, e13725 (2023).

88. Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning. (Springer New York Inc., 2001).

89. Bishop, C. M. Pattern recognition and machine learning. (Springer, 2006).

90. Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction. (MIT Press, 1998).

91. Goodfellow, I., Bengio, Y. & Courville, A. Deep learning. (MIT Press, 2016).

92. Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. Applied linear statistical models. (McGraw-Hill Irwin, 2005).

93. Cortes, C. & Vapnik, V. Support-vector networks. Machine Learning **20**, 273–297 (1995).

94. MacQueen, J. Some methods for classification and analysis of multivariate observations. in vol. 5.1 281–298 (University of California Press, 1967).

95. McCullagh, P. & Nelder, J. A. Generalized linear models. (Chapman & Hall / CRC, 1989).

96. Quinlan, J. R. Induction of decision trees. Machine Learning **1**, 81–106 (1986).

97. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 785–794 (ACM, 2016).

98. Graves, A. & Schmidhuber, J. International Joint Conference on Neural Networks 2005. in vol. 4 2047–2052 (IEEE, 2005).

99. Palotti, J. et al. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. npj Digit. Med. **2**, 1–9 (2019).

100. Sundararajan, K. et al. Sleep classification from wrist-worn accelerometer data using random forests. Sci Rep **11**, 24 (2021).

101. Lutsey, P. L. et al. Objectively measured sleep characteristics and prevalence of coronary artery calcification: The multi-ethnic study of atherosclerosis sleep study. Thorax **70**, 880–887 (2015).

102. Migueles, J. H., Rowlands, A. V., Huber, F., Sabia, S. & Hees, V. T. van. GGIR: A Research Community–Driven Open Source R Package for Generating Physical Activity and Sleep Outcomes From Multi-Day Raw Accelerometer Data. Journal for the Measurement of Physical Behaviour **2**, 188–196 (2019).

103. Trevenen, M. L., Turlach, B. A., Eastwood, P. R., Straker, L. M. & Murray, K. Using hidden Markov models with raw, triaxial wrist accelerometry data to determine sleep stages. Australian & New Zealand Journal of Statistics **61**, 273–298 (2019).

104. Hees, V. T. van et al. Estimating sleep parameters using an accelerometer without sleep diary. Sci Rep **8**, 12975 (2018).

105. Choi, L., Liu, Z., Matthews, C. E. & Buchowski, M. S. Validation of accelerometer wear and nonwear time classification algorithm. Med Sci Sports Exerc **43**, 357–364 (2011).

106. Winkler, E. A. H. et al. Identifying adults' valid waking wear time by automated estimation in activPAL data collected with a 24 h wear protocol. Physiol. Meas. **37**, 1653 (2016).

107. Matthews, C. E., Ainsworth, B. E., Thompson, R. W. & Bassett, D. R. J. Sources of variance in daily physical activity levels as measured by an accelerometer. Medicine & Science in Sports & Exercise **34**, 1376 (2002).

108. King, W. C., Li, J., Leishear, K., Mitchell, J. E. & Belle, S. H. Determining Activity Monitor Wear Time: An Influential Decision Rule. Journal of Physical Activity and Health **8**, 566–580 (2011).

109. Ainsworth, B. E. et al. Recommendations to improve the accuracy of estimates of physical activity derived from self report. J Phys Act Health **9 Suppl 1**, S76–84 (2012).

110. Hecht, A., Ma, S., Porszasz, J., Casaburi, R. & COPD Clinical Research Network. Methodology for using long-term accelerometry monitoring to describe daily activity patterns in COPD. COPD **6**, 121–129 (2009).

111. Ruiz, J. R. et al. Objectively measured physical activity and sedentary time in european adolescents: The HELENA study. Am J Epidemiol **174**, 173–184 (2011).

112. Troiano, R. P., Stamatakis, E. & Bull, F. C. How can global physical activity surveillance adapt to evolving physical activity guidelines? Needs, challenges and future directions. Br J Sports Med **54**, 1468–1473 (2020).

113. Aadland, E., Andersen, L. B., Anderssen, S. A. & Resaland, G. K. A comparison of 10 accelerometer non-wear time criteria and logbooks in children. BMC Public Health **18**, 323 (2018).

114. Toftager, M. et al. Accelerometer data reduction in adolescents: Effects on sample retention and bias. International Journal of Behavioral Nutrition and Physical Activity **10**, 140 (2013).

115. Duncan, S. et al. Wear-time compliance with a dual-accelerometer system for capturing 24-h behavioural profiles in children and adults. Int J Environ Res Public Health **15**, 1296 (2018).

116. Rasmussen, M. G. B. et al. Short-term efficacy of reducing screen media use on physical activity, sleep, and physiological stress in families with children aged 4–14: Study protocol for the SCREENS randomized controlled trial. BMC Public Health **20**, 380 (2020).

117. Zhou, S.-M. et al. Classification of accelerometer wear and non-wear events in seconds for monitoring free-living physical activity. BMJ Open **5**, e007447 (2015).

118. Lee, J. A. & Gill, J. Missing value imputation for physical activity data measured by accelerometer. Stat Methods Med Res **27**, 490–506 (2018).

119. Syed, S., Morseth, B., Hopstock, L. A. & Horsch, A. A novel algorithm to detect non-wear time from raw accelerometer data using deep convolutional neural networks. Sci Rep **11**, 8832 (2021).

120. Carlson, J. A. et al. Validity of two awake wear-time classification algorithms for activPAL in youth, adults, and older adults. Journal for the Measurement of Physical Behaviour **4**, 151–162 (2021).

121. Inan-Eroglu, E. et al. Comparison of a thigh-worn accelerometer algorithm with diary estimates of time in bed and time asleep: The 1970 british cohort study. Journal for the Measurement of Physical Behaviour **4**, 60–67 (2021).

122. Berg, J. D. van der et al. Identifying waking time in 24-h accelerometry data in adults using an automated algorithm. Journal of Sports Sciences **34**, 1867–1873 (2016).

123. Doherty, A. et al. Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. PLOS ONE **12**, e0169649 (2017).

124. Silva, I. C. da et al. Physical activity levels in three Brazilian birth cohorts as assessed with raw triaxial wrist accelerometry. International Journal of Epidemiology **43**, 1959–1968 (2014).

125. Anderson, K. N. et al. Assessment of sleep and circadian rhythm disorders in the very old: The newcastle 85+ cohort study. Age and Ageing **43**, 57–63 (2014).

126. Lockley, S. W., Skene, D. J. & Arendt, J. Comparison between subjective and actigraphic measurement of sleep and sleep rhythms. Journal of Sleep Research **8**, 175–183 (1999).

127. Skovgaard, E. L., Pedersen, J., Møller, N. C., Grøntved, A. & Brønd, J. C. Manual annotation of time in bed using free-living recordings of accelerometry data. Sensors (Basel) **21**, 8442 (2021).

128. Rasmussen, M. G. B. et al. Feasibility of two screen media reduction interventions: Results from the SCREENS pilot trial. PLOS ONE **16**, e0259657 (2021).

129. Jaeschke, L., Steinbrecher, A., Jeran, S., Konigorski, S. & Pischon, T. Variability and reliability study of overall physical activity and activity intensity levels using 24 h-accelerometry-assessed data. BMC Public Health **18**, 530 (2018).

130. Audacity Team. Audacity® software is copyright © 1999–2021 audacity team. (2021).

131. Visplore – software for visual time series analysis.

132. Open Source Data Labeling.

133. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin **86**, 420–428 (1979).

134. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. Psychological Methods **1**, 30–46 (1996).

135. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med **15**, 155–163 (2016).

136. William Revelle. Psych: Procedures for psychological, psychometric, and personality research. (Northwestern University, 2023).

137. Bland, J. M. & Altman, D. G. Measuring agreement in method comparison studies. Stat Methods Med Res **8**, 135–160 (1999).

138. Aili, K., Åström-Paulsson, S., Stoetzer, U., Svartengren, M. & Hillert, L. Reliability of actigraphy and subjective sleep measurements in adults: The design of sleep assessments. J Clin Sleep Med **13**, 39–47 (2017).

139. Haghayegh, S., Khoshnevis, S., Smolensky, M. H. & Diller, K. R. Application of deep learning to improve sleep scoring of wrist actigraphy. Sleep Med **74**, 235–241 (2020).

140. Yavuz-Kodat, E. et al. Validity of Actigraphy Compared to Polysomnography for Sleep Assessment in Children With Autism Spectrum Disorder. Frontiers in Psychiatry **10**, 551 (2019).

141. Skovgaard, E. L. et al. Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings. Sci Rep **13**, 2496 (2023).

142. Troiano, R. P. et al. Physical activity in the united states measured by accelerometer. Med Sci Sports Exerc **40**, 181–188 (2008).

143. Hees, V. T. van et al. Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. PLoS One **6**, e22922 (2011).

144. Hees, V. T. van et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. PLOS ONE **8**, e61691 (2013).

145. Syed, S., Morseth, B., Hopstock, L. A. & Horsch, A. Evaluating the performance of raw and epoch non-wear algorithms using multiple accelerometers and electrocardiogram recordings. Sci Rep **10**, 5866 (2020).

146. Pedersen, N. H. et al. Protocol for evaluating the impact of a national school policy on physical activity levels in danish children and adolescents: The PHASAR study - a natural experiment. BMC Public Health **18**, 1245 (2018).

147. Kuhn, M. & Wickham, H. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. (2020).

148. Therneau, T. & Atkinson, B. Rpart: Recursive partitioning and regression trees. (2022).

149. Hutto, B. et al. Identifying accelerometer nonwear and wear time in older adults. International Journal of Behavioral Nutrition and Physical Activity **10**, 120 (2013).

150. Cooper, A. R. et al. Objectively measured physical activity and sedentary time in youth: The international children's accelerometry database (ICAD). International Journal of Behavioral Nutrition and Physical Activity **12**, 113 (2015).

151. Barouni, A. et al. Ambulatory sleep scoring using accelerometers-distinguishing between nonwear and sleep/wake states. PeerJ **8**, e8284 (2020).

152. Mn, A., N, N., R, S., L, W. & Sg, T. Non-wear or sleep? Evaluation of five non-wear detection algorithms for raw accelerometer data. Journal of sports sciences **38**, (2020).

153. Vert, A. et al. Detecting accelerometer non-wear periods using change in acceleration combined with rate-of-change in temperature. BMC Medical Research Methodology **22**, 147 (2022).

154. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol **69**, 245–247 (2016).

155. Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. M. Prognosis and prognostic research: Validating a prognostic model. BMJ **338**, b605 (2009).

156. Sagelv, E. H. et al. Physical activity levels in adults and elderly from triaxial and uniaxial accelerometry. The Tromsø Study. PLOS ONE **14**, e0225670 (2019).

157. Pedersen, J. et al. Effects of Limiting Recreational Screen Media Use on Physical Activity and Sleep in Families With Children: A Cluster Randomized Clinical Trial. JAMA pediatrics **176**, 741–749 (2022).

158. Walch, O., Huang, Y., Forger, D. & Goldstein, C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep **42**, zsz180 (2019).

159. Lewy, A. J., Wehr, T. A., Goodwin, F. K., Newsome, D. A. & Markey, S. P. Light suppresses melatonin secretion in humans. Science (New York, N.Y.) **210**, 1267–1269 (1980).

160. Galland, B. C. et al. Establishing normal values for pediatric nighttime sleep measured by actigraphy: a systematic review and meta-analysis. Sleep **41**, (2018).

161. Ohayon, M. et al. National Sleep Foundation's sleep quality recommendations: first report. Sleep Health **3**, 6–19 (2017).

162. Galland, B. C., Taylor, B. J., Elder, D. E. & Herbison, P. Normal sleep patterns in infants and children: A systematic review of observational studies. Sleep Med Rev **16**, 213–222 (2012).

163. Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Computation **9**, 1735–1780 (1997).

164. Sano, A., Chen, W., Lopez-Martinez, D., Taylor, S. & Picard, R. W. Multimodal ambulatory sleep detection using LSTM recurrent neural networks. IEEE J Biomed Health Inform **23**, 1607–1617 (2019).

165. Chen, Z., Wu, M., Cui, W., Liu, C. & Li, X. An attention based CNN-LSTM approach for sleep-wake detection with heterogeneous sensors. IEEE J Biomed Health Inform **25**, 3270–3277 (2021).

166. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software **33**, 1–22 (2010).

167. Venables, W. N. & Ripley, B. D. Modern applied statistics with s. (Springer, 2002).

168. Chen, T. et al. Xgboost: Extreme gradient boosting. (2023).

169. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002).

170. Hvitfeldt, E. Themis: Extra recipes steps for dealing with unbalanced data. (2023).

171. Kushida, C. A. et al. Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. Sleep Med **2**, 389–396 (2001).

172. DiCiccio, T. J. & Efron, B. Bootstrap confidence intervals. Statistical Science **11**, 189–228 (1996).

173. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, 2023).

174. Wickham, H. et al. Welcome to the tidyverse. Journal of Open Source Software **4**, 1686 (2019).

175. Van Rossum, G. & Drake, F. L. Python 3 reference manual. (CreateSpace, 2009).

176. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. in 80248035 (Curran Associates, Inc., 2019).

177. Patterson, M. R. et al. 40 years of actigraphy in sleep medicine and current state of the art algorithms. npj Digit. Med. **6**, 1–7 (2023).

178. Dozy, D. & Slumberton, S. Snooze analytics: A comprehensive guide to overthinking bedtime. Journal of Procrastination and Pillow Fluffing **69**, Zzz–Zzzz (2023).

179. Plekhanova, T. et al. Validation of an automated sleep detection algorithm using data from multiple accelerometer brands. J Sleep Res **32**, e13760 (2023).

180. Hees, V. van, Charman, S. & Anderson, K. Newcastle polysomnography and accelerometer data.

181. SOMNOtouch™ RESP - home sleep testing (HST).

182. Hees, V. T. van et al. Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: An evaluation on four continents. Journal of Applied Physiology **117**, 738–744 (2014).

183. Burazeri, G., Gofin, J. & Kark, J. D. Siesta and mortality in a Mediterranean population: a community study in Jerusalem. Sleep **26**, 578–584 (2003).

184. Makino, S. et al. Association between nighttime sleep duration, midday naps, and glycemic levels in Japanese patients with type 2 diabetes. Sleep Medicine **44**, 4–11 (2018).

185. Pastor-Pellicer, J., Zamora-Martínez, F., España-Boquera, S. & Castro-Bleda, M. J. F-Measure as the Error Function to Train Neural Networks. in (eds. Rojas, I., Joya, G. & Gabestany, J.) 376–384 (Springer, 2013).

186. Oyebode, O., Fowles, J., Steeves, D. & Orji, R. Machine learning techniques in adaptive and personalized systems for health and wellness. International Journal of Human–Computer Interaction **39**, 1938–1962 (2023).

187. Mukherjee, S. et al. An official american thoracic society statement: The importance of healthy sleep. Recommendations and future priorities. American Journal of Respiratory and Critical Care Medicine **191**, 1450–1458 (2015).

188. Quante, M. et al. Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms. Nature and Science of Sleep **10**, 13–20 (2018).

189. Cassidy, S., Chau, J. Y., Catt, M., Bauman, A. & Trenell, M. I. Cross-sectional study of diet, physical activity, television viewing and sleep duration in 233,110 adults from the UK Biobank; the behavioural phenotype of cardiovascular disease and type 2 diabetes. BMJ open **6**, e010038 (2016).

190. Espiritu, J. R. D. Aging-related sleep changes. Clinics in Geriatric Medicine **24**, 1–14, v (2008).

191. Altaf, Q. A. et al. Obstructive Sleep Apnea and Retinopathy in Patients with Type 2 Diabetes. A Longitudinal Study. American Journal of Respiratory and Critical Care Medicine **196**, 892–900 (2017).

192. Heinzer, R. et al. Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study. The Lancet. Respiratory Medicine **3**, 310–318 (2015).

193. Zhang, G.-Q. et al. The National Sleep Research Resource: towards a sleep data commons. Journal of the American Medical Informatics Association: JAMIA **25**, 1351–1358 (2018).

194. Chen, X. et al. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). Sleep **38**, 877–888 (2015).

195. Zambotti, M. de, Cellini, N., Goldstone, A., Colrain, I. M. & Baker, F. C. Wearable Sleep Technology in Clinical and Research Settings. Medicine and Science in Sports and Exercise **51**, 1538–1557 (2019).

196. Rentz, L. E., Ulman, H. K. & Galster, S. M. Deconstructing Commercial Wearable Technology: Contributions toward Accurate and Free-Living Monitoring of Sleep. Sensors **21**, 5071 (2021).

197. Altini, M. & Kinnunen, H. The Promise of Sleep: A Multi-Sensor Approach for Accurate Sleep Stage Detection Using the Oura Ring. Sensors **21**, 4302 (2021).

198. Herzig, D. et al. Reproducibility of heart rate variability is parameter and sleep stage dependent. Frontiers in Physiology **8**, (2018).

199. Chouchou, F. & Desseilles, M. Heart rate variability: A tool to explore the sleeping brain? Frontiers in Neuroscience **8**, (2014).

200. Long, X. et al. Detection of nocturnal slow wave sleep based on cardiorespiratory activity in healthy adults. IEEE Journal of Biomedical and Health Informatics **21**, 123–133 (2017).

201. Muzet, A. et al. Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: Comparison with results based on polysomnography. Sleep Medicine **21**, 47–56 (2016).

List of Appendices

- **Appendix I**: Manual Annotation of Time in Bed Using Free-Living Recordings of Accelerometry Data
- **Appendix II**: Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings
- **Appendix III**: Improving Sleep Quality Estimation in Children and Adolescents: A Comparative Study of Machine Learning and Deep Learning Techniques Utilizing Free-Living Accelerometer Data from Thigh-Worn Devices and EEG-Based Sleep Tracking

Appendix I

Manual Annotation of Time in Bed Using Free-Living Accelerometry Data

DOI: https://doi.org/10.3390/s21248442

sensors

MDPI

Article

Manual Annotation of Time in Bed Using Free-Living Recordings of Accelerometry Data

Esben Lykke Skovgaard *, Jesper Pedersen, Niels Christian Møller, Anders Grøntved and Jan Christian Brønd

Centre of Research in Childhood Health, Research Unit for Exercise Epidemiology, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, 5230 Odense, Denmark; jespedersen@health.sdu.dk (J.P.); ncmoller@health.sdu.dk (N.C.M.); agroentved@health.sdu.dk (A.G.); jbrond@health.sdu.dk (J.C.B.)

* Correspondence: eskovgaard@health.sdu.dk

Abstract: With the emergence of machine learning for the classification of sleep and other human behaviors from accelerometer data, the need for correctly annotated data is higher than ever. We present and evaluate a novel method for the manual annotation of in-bed periods in accelerometer data using the open-source software Audacity®, and we compare the method to the EEG-based sleep monitoring device Zmachine® Insight+ and self-reported sleep diaries. For evaluating the manual annotation method, we calculated the inter- and intra-rater agreement and agreement with Zmachine and sleep diaries using interclass correlation coefficients and Bland–Altman analysis. Our results showed excellent inter- and intra-rater agreement and excellent agreement with Zmachine and sleep diaries. The Bland–Altman limits of agreement were generally around ±30 min for the comparison between the manual annotation and the Zmachine timestamps for the in-bed period. Moreover, the mean bias was minuscule. We conclude that the manual annotation method presented is a viable option for annotating in-bed periods in accelerometer data, which will further qualify datasets without labeling or sleep records.

Keywords: sleep; accelerometry; labeling; machine learning; physical activity; human behavior; circadian rhythms; classification; sleep/wake cycles; annotation; wearable sensors

Citation: Skovgaard, E.L.; Pedersen, J.; Møller, N.C.; Grøntved, A.; Brønd, J.C. Manual Annotation of Time in Bed Using Free-Living Recordings of Accelerometry Data. *Sensors* **2021**, *21*, 8442. https://doi.org/10.3390/s21248442

Academic Editor: Steven Vos

Received: 12 November 2021
Accepted: 14 December 2021
Published: 17 December 2021

1. Introduction

Utilizing machine learning with the identification of sleep, physical activity behavior, or non-wear from accelerometry data provides the ability to model very complex and non-linear relationships, which is not possible with more simple statistical methods, like multiple linear or logistic regression [1]. However, the use of supervised machine learning algorithms demands large amounts of accurate annotated data to provide sufficient accuracy and to ensure generalizability [2]. Sleep is increasingly recognized as a critical component of the healthy development of children and overall health [3–5], and healthy sleep is generally defined by adequate duration, appropriate timing, good quality, and the absence of sleep disturbances or disorders [6].

Nevertheless, only scarce efforts to assess sleep measures from accelerometer data have been successfully approached with advanced machine learning techniques [7] providing researchers with inexpensive and minimally invasive methods. Valid objective measures using accurate automated scoring of sleep and wake time are important to provide valuable insights into the circadian rhythms on an individual and population-wide basis. The gold standard for objective sleep assessment is polysomnography (PSG), which is based on the continuous recording of electroencephalographic (EEG), electromyographic (EMG), and electrooculographic (EOG) activity via surface electrodes.

The assessment of sleep from PSG provides a detailed description of individuals' sleep architecture through the identification of various sleep stages in addition to the more general sleep outcomes, like sleep duration and timing [8]. The PSG method is costly and

burdensome in terms of technician support for sensor application/removal, overnight monitoring (for in-lab PSG), and manual record scoring, in addition to being intrusive for the patient due to the necessity of wearing multiple sensors on the scalp and face throughout the night.

Recent studies have attempted to identify PSG-assessed sleep-wake classification with wrist acceleration using machine-learning techniques [7,9,10]. One study by Sundararajan et al. [9] attempted this using a random forest machine learning algorithm. The results from the study revealed an F1 score of 73.9% as an estimate of overall accuracy for the identification of sleep-wake stages. Moreover, the study showed a large false discovery rate (i.e., the true wake time was predicted as sleep by the algorithm) for the identification of sleep.

This may primarily be a result of the intrinsic limitations of a wrist-worn accelerometer incorrectly classifying quiet wakefulness as sleep and, secondly, may be explained by the number of participants and the use of single-night PSG-recordings restricting information on inter- and intraindividual variation. Thus, the poor false discovery rate could potentially be balanced out by increasing the number of subjects in the study and/or by increasing the number of consecutive days of recordings for more information on the variation in the movement behavior of the subjects during sleep hours.

Accelerometry provides researchers with an inexpensive and minimally invasive method, which has the potential to play an important role in the assessment of sleep duration and timing characterization since it is more practical and suitable than PSG for prolonged recordings (i.e., multiple consecutive days) in non-laboratory settings [11]. However, the limitations of accelerometry must be acknowledged, which is also emphasized by the poor results for the identification of sleep vs. wakefulness with wrist acceleration presented with the study by Sundararajan et al.

Thus, accelerometry does impose certain limitations as a methodology to identifying subject's sleep behavior and developing new algorithms should not focus on sleep staging but rather the sleep timing (bedtime/wake-up time) and specifically the sleep/wake state of the subject. Moreover, the accurate identification of the sleep-wake state from accelerometry is most optimally approached with accelerometry recordings covering at least 7–10 days of measurement for each subject to ensure the appropriate day-to-day variation and movement behavior during sleep hours.

Developing supervised machine learning algorithms to identify the sleep/wake cycle from multiple days of accelerometry recordings of individuals requires the annotation of the data to identify time in bed and specifically when the participants go to bed and when they get out of bed. Although there is an obvious distinction between time in bed and actual sleep time, accelerometry as a surrogate measure of sleep is widely used in the literature [12–15] due to the many practical advantages of using accelerometry compared to more intricate methodologies for the detection of sleep.

The time in bed annotation could be established from individual sleep diaries, EEG-based recordings [16], systems for the recording of tracheal sounds [17,18], etc.; however, such additional data are not recorded in conjunction with accelerometry within many studies, which leaves a substantial data resource ready for enrichment. If individual time in bed periods can be accurately annotated without additional data for correct labeling, it would provide the option to use the accelerometer data to facilitate the improvement of existing algorithms or the development of new supervised machine learning algorithms. Currently, there are no accurate or easy-to-use methods for the manual annotation of time in bed from accelerometry.

The aims of the present study were to (1) describe a method for the manual annotation of subjects' individual bedtime and time out of bed with raw unprocessed accelerometry, (2) evaluate the accuracy of the manual annotation to predict time in bed/out of bed obtained using a single channel EEG-based sleep staging system and a sleep diary, and (3) to evaluate the inter- and intra-rater reliability of annotations.

2. Materials and Methods

2.1. Study Population

Data for the current study originates from the SCREENS pilot trial (www.clinicaltrials.gov (accessed on 15 May 2020), NCT03788525), which is a two-arm parallel-group cluster-randomized trial with two intervention groups and no control group [19,20]. Data were collected between October 2018 and March 2019.

The collection of data was reported to the local data protection department SDU RIO (ID: 10.391) in agreement with the rules of the Danish Data Protection Agency.

Families in the municipality of Middelfart in Denmark were invited to participate if they had at least one child aged 6–10 years residing in the household (n = 1686). Based on survey responses, families were eligible to participate if the contacted parent's total screen media use was above the median amount (2.7 h/day) based on all respondents (n = 394) and if all children in the household were older than 3.9 years.

The latter was to avoid potential disturbances of sleep measurement due to an infant or toddler's pattern of nocturnal awakening. For further details on inclusion and exclusion criteria see Pedersen et al. [21] In total, data from 14 children and 19 adults were included in the present study. The included participants were not instructed to change their sleep and bedtime behavior as a part of the interventions. The napping behavior, if any, of the participants was irrelevant to the current study as we focused on their nightly sleep time monitored by the EEG-based sleep staging system.

2.2. Actigraphy

Both adults and children underwent 24-h accelerometry recording using two Axivity AX3 (Axivity Ltd., Newcastle upon Tyne, UK) triaxial accelerometers. The Axivity AX3 is a small (dimensions: 23 mm × 32.5 mm × 7.6 mm) weighing only 11 g. Sensitivity was set to ±8 g and the sampling frequency to 50 Hz. The accelerometers were worn at two anatomical locations; one fixated to the body in a pocket attached to a belt worn around the waist, where the sensor was placed on the right hip with the USB connector facing away from the right side of the body.

A second belt was worn around the right thigh midway between the hip and the knee, where the accelerometer was placed in a pocket with the USB connector facing away from the body. The devices were worn for 1 week (seven consecutive days) at baseline and at follow-up, which corresponds to the recommended number of days required to reliably estimate habitual physical activity [22].

2.3. Zmachine® Insight+

Concurrent with the accelerometer recordings, both adults and children underwent sleep assessments for 3–4 nights at baseline and 3 nights at follow-up using the Zmachine® (ZM) Insight+ model DT-200 (General Sleep Corporation, Cleveland, OH, USA), Firmware version 5.1.0) concurrently with the actigraphy recording. The device measures sleep by single-channel EEG from the differential mastoid (A1–A2) EEG location on a 30-s epoch basis. The sleep apparatus is developed for use in a free-living setting for objective measurement of sleep, including measurement of sleep duration and sleep stage classification, as well as computation of sleep-specific quantities, e.g., latency to the respective sleep stages.

The algorithm in ZM has been compared to polysomnography (PSG) in adults with and without chronic sleep issues within a laboratory setting [23,24], and we found that ZM can be feasibly applied to children and adults for multiple days of measurements in free-living [12]. The ZM device demonstrated a high degree of validity for detecting sleep versus awake with a sensitivity, specificity, positive predictive value, and negative predictive values of 95.5%, 92.5%, 98%, and 84.2%, respectively [24].

Three electrodes (Ambu A/S, Ballerup, Denmark, type: N-00-S/25) are mounted on the mastoids (signal) and the back of the neck (ground). Thirty minutes before the participants plan to go to bed to sleep, the skin areas are cleansed with an alcohol swab,

and then electrodes are attached to the skin. An EEG cable connects the three electrodes to the ZM device, where a sensor check is performed to detect whether one or more electrodes are not mounted correctly. If there are sensor problems, these are solved swiftly by a simple change of said electrodes. Participants also reported their bedtimes and time of awakening each day using a prospective daily diary. Parents reported the bedtimes and times of awakening on behalf of their child.

2.4. Audacity

Audacity is a free software tool that was developed for audio editing. The software project was originally started by Dominic Mazzoni and Roger Dannenberg in the fall of 1999 as part of a research project at Carnegie Mellon University. Audacity was initially released as an open-source audio editor in May 2000. Since then, the software has been community-developed adding hundreds of features, providing complete support for professional-quality 24-bit and 32-bit audio, a complete Manual and support for many different languages, and millions of copies have been distributed. Today, Audacity is being maintained by a team of volunteers located in many different countries. Audacity is distributed under the terms of the GNU General Public License. Everyone is free to use this application for any personal, educational, or commercial purpose.

In the context of analyzing accelerometer data, Audacity is especially useful as it enables the researchers to effortlessly inspect high-resolution raw accelerometer data with a high degree of precision. It is possible to rapidly zoom in to inspect portions of the accelerometer recording in detail (e.g., to inspect certain behavior around bedtime) or zoom out to get an overview (e.g., a whole week). Moreover, Audacity provides a high-resolution labeling function that can be used for the annotation of the accelerometry data. All labels created can be stored in a separate file and subsequently used in the ML algorithm. The ability to manually inspect high-resolution raw accelerometer data at the level of detail provided by Audacity is, to the knowledge of the authors of the current study, unprecedented in other software.

In the Audacity software, it is possible to combine more than 100 channels of data, which provides the ability to combine different signal features derived from the acceleration. Adding multiple signal features together provides an interesting option that might facilitate the visual interpretation and classification of the underlying behavior. However, adding too many signal features could have negative consequences for the accurate identification of the behavior of interest. We combined a total of seven independent signal features. The classification of "lying" within the first feature is derived as follows: the inclination of the hip accelerometer exceeds 65 degrees, and the thigh accelerometer is concurrently classified as "sitting" by the activity type classification algorithm by Skotte et al. [25]. The remaining signal features, excluding "time", were selected from algorithm by Skotte et al. directly. All included features are summarized and described in Table 1. The accelerometry signal features are described with respect to the longitudinal axis of the body. The features generated from the accelerometry data is processed using a window length of two seconds (60 samples) and a 50% overlap (30 samples), providing a resolution of one second. The algorithm by Skotte et al. and the algorithm producing the first feature are solely based on the inclination(s) of the accelerometer(s) and, as such, can be used not only to assess the time in bed but also the posture of the participants. Therefore, this is not a precise indicator of the to-bed/out-of-bed timestamps.

Figures 1 and 2 shows examples of the visual Audacity interface with all seven signal features as listed in Table 1. Figure 1 is a seven-day overview, and Figure 2 presents a zoomed view of approximately 24 h and with a single annotated night.

Table 1. Signal features for the detection of in-bed periods in Audacity.

Name	Description	Values	Visual Interpretation
Lying	The classification of lying based on the thigh and back	1: lying −1: not lying	Lying position
Activity	The classification of activity type	1: Standing, moving, or walking 0: Sitting −1: Other activity	This feature will guide the rater to identify periods of activity prior to correct bedtime
Time	Time categorized into four-hour windows	−1: 00:00–04:00 and so on throughout the 24 h cycle	Time of day
Thigh-SDacc	Standard deviation of the acceleration on the longitudinal axis of the thigh	−1: No movement	Proportion of leg movement
Thigh-Inclination	Inclination angle of the thigh device in relation to the longitudinal axis of the thigh	The −1 to 1 range represents the −180 to 180 degrees inclination angle	Inclination angle of the thigh
Hip-SDacc	Standard deviation of the hip acceleration on the longitudinal axis of the torso	−1: No movement	Proportion of whole-body movement
Hip-Inclination	Inclination of the hip device in relation to the longitudinal axis of the torso	The −1 to 1 range represents the −180 to 180 degrees inclination angle	Inclination angle of the body/hip

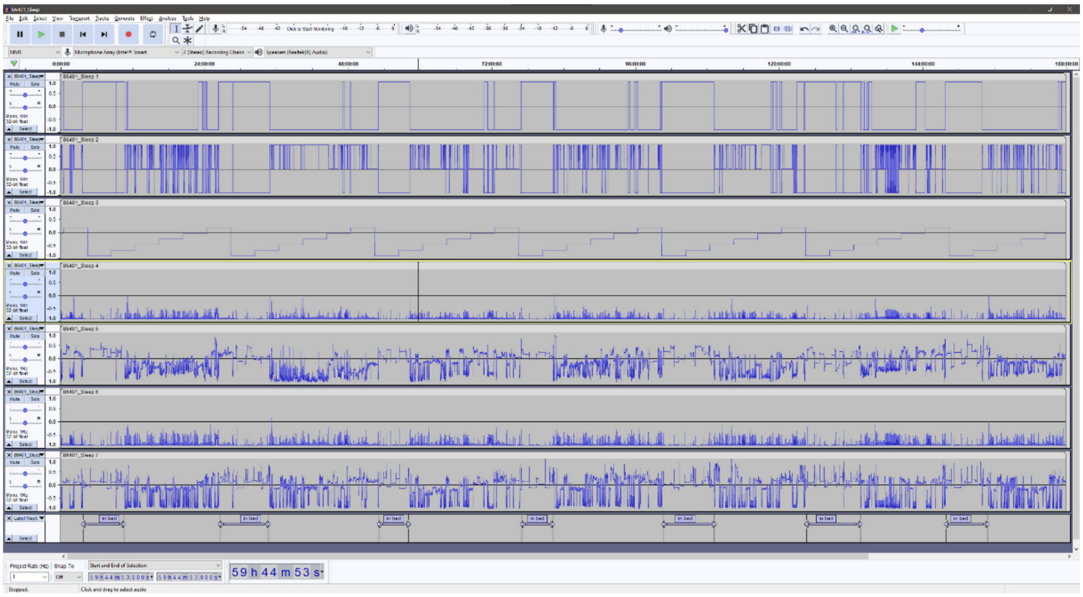

Figure 1. Screenshot of the Audacity interface showing the seven horizontal panels representing the included signal features. See Table 1 for a detailed description of the features.

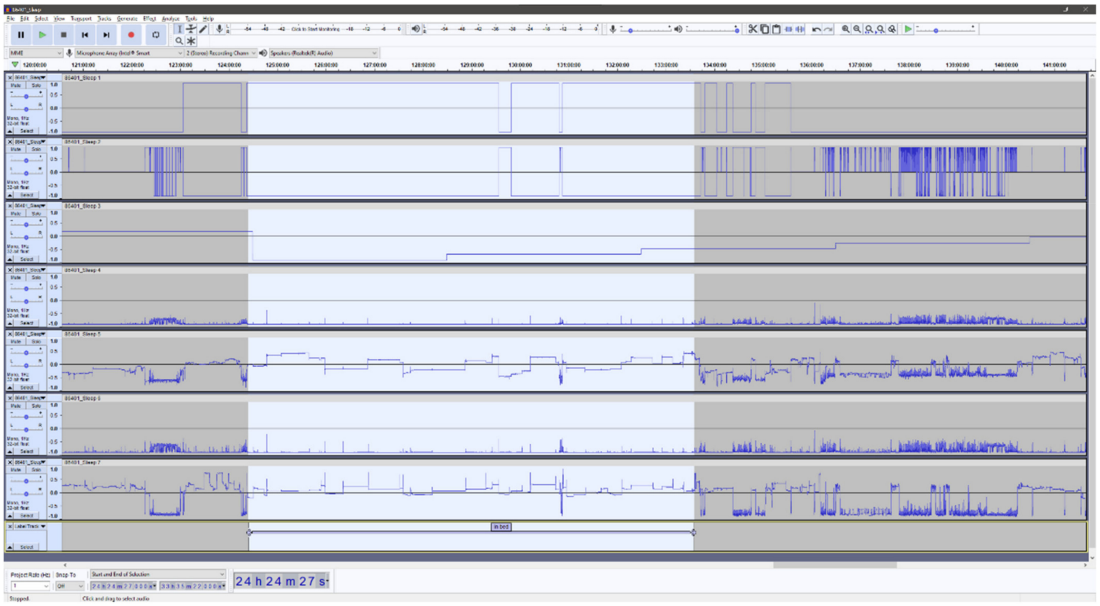

Figure 2. Screenshot of the Audacity interface when zoomed in on a single night for the labeling of the in-bed period. The seven horizontal panels represent the included signal features. See Table 1 for a detailed description of features.

2.5. Manual Annotation by the Raters

The three raters were all researchers who had prior experience working with accelerometer data, and thus, had some understanding and knowledge on how to interpret the different data channels available. The raters labeled each wav-file independently of each other with in-bed and out-of-bed timestamps and exported the corresponding labels as text files. Each file was labeled twice (round 1 and round 2) for test–retest purposes. The raters were at no time aware of previous annotations made by themselves or by the other raters.

2.6. Ground Truth

The ZM ground truth labels of the time in-bed and out-of-bed events were derived from the sleep staging data of the ZM as the first and last non-sensor-problem event for the night. If the ZM reported the beginning or the end of the recording as having sensor problems, the corresponding night was discarded from further analysis. Sensor problems most commonly occur due to poor attachment of the electrodes. All subjects were instructed to attach the ZM and turn it on at the exact time when the participants went to bed and to remove upon awakening. The timestamps of these events were used as the ground truth values.

2.7. Statistical Analysis

All statistical analyses were performed using R statistical (R Core Team, Vienna, Austria) software version 4.0.2 (22 June 2020), RStudio (RStudio Inc., Boston, MA, USA) version 1.1.456. Descriptive characteristics were computed using medians and interquartile ranges for continuous variables and proportions for categorical variables. Characteristics are presented separately for children and adults.

Agreement analyses were performed using intraclass correlation coefficient (ICC) and Bland–Altman analysis. Furthermore, to illustrate the overall agreement and symmetry of methods, probability density distribution plots are shown. The ICC is an index that, con-

trary to Pearson correlation, assesses not only how well correlated the two techniques are but also if they are equal. An ICC < 0.5 indicates poor agreement, 0.5 < ICC > 0.75 indicates moderate agreement, 0.75 < ICC > 0.9 indicates good agreement, and ICC > 0.90 indicates excellent agreement [26].

In the current study, interpretations of the ICCs are based on the corresponding 95% confidence intervals in accordance with guidelines [26]. Bland–Altman analyses allow examining the degree of agreement between two measurement techniques [27]. The mean of the differences between two techniques (representing the mean bias) and limits of agreement (which are defined as a deviation from the mean superior to two standard deviations) are calculated. A positive bias/mean difference indicates an underestimation (earlier) of the to-bed or out-of-bed timestamp, while a negative difference indicates an overestimation (later) compared to ZM.

3. Results

Descriptive characteristics of the included subjects of the current study are reported in Table 2.

Table 2. Descriptive characteristics of the study participants.

Population (*N* = 33)	
Children	
N	14
Gender (% female)	28.6
Age (years)	9 (7–10)
Adults	
N	19
Gender (% female)	57.9
Age (years)	42 (39–46)
ISCED	
0–3 (%)	36.8
4–6 (%)	47.4
7–8 (%)	15.8

ISCED, International Standard Classification of Education

3.1. Intraclass Correlation Coefficient Analyses

The ICC analyses highlighted an excellent agreement between ZM and manual in-bed annotation for time to bed and time out of bed across both rounds 1 and 2 and at the baseline and follow-up with the lower limits of the confidence intervals all above 0.9 (see Table 3).

Table 3. Intraclass correlation coefficients between ZM and the average of the manual annotations between the three raters.

	Baseline (*n* = 94 Nights)		**Follow-Up (*n* = 54 Nights)**	
	Round 1 ICC (95% CI)	Round 2 ICC (95% CI)	Round 1 ICC (95% CI)	Round 2 ICC (95% CI)
To bed	0.98 (0.98; 0.99)	0.98 (0.96; 0.98)	0.96 (0.94; 0.98)	0.95 (0.92; 0.97)
Out of bed	0.98 (0.97; 0.99)	0.98 (0.96; 0.98)	0.98 (0.97; 0.99)	0.97 (0.95; 0.98)

Round 1 and round 2 refers to the first and second round of annotation.

Excellent agreement was also observed between self-report and ZM for both baseline data and follow-up data, which is indicated by the lower limit of the 95% confidence interval has values no less than 0.94 (see Table 4).

Table 4. Intraclass correlation coefficients between self-report and ZM.

	Baseline (*n* = 94 Nights)	**Follow-Up (*n* = 54 Nights)**
	ICC (95% CI)	ICC (95% CI)
To bed	0.98 (0.98; 0.99)	0.96 (0.94; 0.98)
Out of bed	0.98 (0.97; 0.99)	0.98 (0.96; 0.99)

The ICCs of the agreement between the three manual raters' ability to annotate the to bed and out of bed timestamps showed good to excellent agreement as seen by the lower limits of the 95% confidence intervals no less than 0.88. A slight tendency of difference of the ICCs can be seen when comparing the to-bed to the out-of-bed timestamps (see Table 5).

Table 5. Intraclass correlation coefficients between manual raters.

	Baseline (*n* = 110 Nights)		**Follow-Up (*n* = 62 Nights)**	
	Round 1 ICC (95% CI)	Round 2 ICC (95% CI)	Round 1 ICC (95% CI)	Round 2 ICC (95% CI)
To bed	0.91 (0.88; 0.94)	0.92 (0.89; 0.94)	0.94 (0.9; 0.96)	0.97 (0.95; 0.98)
Out of bed	0.93 (0.9; 0.95)	0.97 (0.96; 0.98)	0.97 (0.96; 0.98)	0.98 (0.98; 0.99)

Round 1 and round 2 refers to the first and second round of annotation.

The ICCs for the test–retest reliability showed good to excellent agreement for each rater between rounds 1 and 2 for both baseline data and follow-up data (see Table 6). This is seen by the lower limits of the 95% confidence intervals values of no less than 0.86. Although the ICCs are similar, it seems that raters 1 and 3 showed lower agreement when annotating the baseline to-bed timestamp compared to the later annotations, whereas the ICC scores of rater 2 did not elicit this behavior.

Table 6. Test–retest intraclass correlation coefficients between the first and second round of manual annotations.

	Baseline (*n* = 110 Nights)		**Follow-Up (*n* = 62 Nights)**	
	To Bed ICC (95% CI)	Out of Bed ICC (95% CI)	To Bed ICC (95% CI)	Out of Bed ICC (95% CI)
Rater 1	0.91 (0.87; 0.94)	0.98 (0.98; 0.99)	0.96 (0.94; 0.98)	1.00 (0.99; 1.00)
Rater 2	0.97 (0.96; 0.98)	0.91 (0.87; 0.94)	0.91 (0.86; 0.95)	0.99 (0.98; 0.99)
Rater 3	0.91 (0.87; 0.94)	0.96 (0.94; 0.97)	0.98 (0.97; 0.99)	0.98 (0.97; 0.99)

3.2. Bland–Altman Analyses

The bias and upper and lower limits of agreement with corresponding confidence intervals for the comparison of the manual annotation and self-report in relation to ZM are presented in Table 7. Biases for the manual annotation compared to ZM is in the range of -6 min to 5 min, while self-report produced slightly lower biases in comparison to ZM. Generally, the limits of agreement seem to be of the same magnitude regardless of the method comparison.

Table 7. Bland–Altman analysis of inter-method agreement between manual annotation and ZM as well as self-report and ZM. All estimates are in minutes.

Method	Bias (95% CI)	Upper LOA (95% CI)	Lower LOA (95% CI)
Baseline, to bed (n = 94)			
Manual, round 1	3.02 (−0.44; 6.47)	−30.04 (−35.96; −24.12)	36.07 (30.15; 42)
Manual, round 2	0.48 (−2.42; 3.39)	−27.3 (−32.28; −22.32)	28.27 (23.29; 33.24)
Self-report	1.23 (−1.57; 4.03)	−25.56 (−30.37; −20.76)	28.02 (23.21; 32.82)
Baseline, out of bed (n = 94)			
Manual, round 1	0.53 (−2.34; 3.4)	−26.9 (−31.82; −21.99)	27.96 (23.05; 32.88)
Manual, round 2	0.98 (−1.47; 3.43)	−22.49 (−26.7; −18.28)	24.45 (20.24; 28.66)
Self-report	−2.79 (−5.26; −0.32)	−26.45 (−30.69; −22.21)	20.87 (16.63; 25.11)
Follow-up, to bed (n = 54)			
Manual, round 1	−6.08 (−11.34; −0.83)	−43.81 (−52.84; −34.77)	31.64 (22.61; 40.67)
Manual, round 2	−0.4 (−5.3; 4.51)	−35.6 (−44.03; −27.17)	34.8 (26.37; 43.23)
Self-report	0.77 (−4.08; 5.62)	−34.06 (−42.4; −25.72)	35.59 (27.25; 43.93)
Follow-up, out of bed (n = 54)			
Manual, round 1	4.95 (0.65; 9.25)	−25.95 (−33.35; −18.55)	35.85 (28.45; 43.25)
Manual, round 2	2.57 (−0.76; 5.89)	−21.3 (−27.02; −15.59)	26.44 (20.72; 32.15)
Self-report	0.56 (−3.62; 4.74)	−29.45 (−36.64; −22.26)	30.57 (23.39; 37.76)

3.3. Density Plots

The probability density distribution for the difference between the to-bed and out-of-bed scoring for the manual annotation and self-report compared to ZM is shown in Figure 3. These plots function as a visual representation of the bias and spread around zero of the manual annotations and self-report in comparison to ZM as seen previously [10].

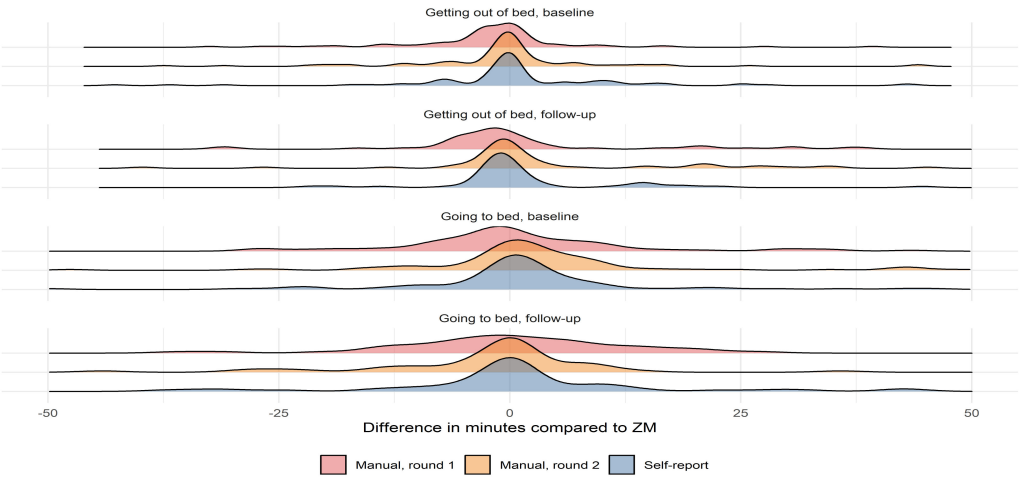

Figure 3. Probability density distributions for differences between manual in-bed annotations and self-report compared to ZM.

4. Discussion

This is the first study to describe a method for the manual annotation of in-bed periods with accelerometry data and to evaluate the accuracy of the method with multiple raters and compared to sleep assessed with EEG -based methodology Zmachine Insight+® and self-reported sleep as reference methodologies. When all interpretations of the ICC analyses were based on the lower limit of the 95% confidence interval, our results showed (1) good-to-excellent interrater reliability, (2) the test–retest reliability (or intra-rater reliability)

showed good to excellent agreement for all three raters between their first and second round of in-bed annotations, (3) compared to ZM, the average of the manual in-bed annotation method for all three raters showed agreements ranging from good to excellent, and (4) the agreement between the self-reported in-bed timestamps and ZM were good to excellent. Furthermore, the Bland–Altman analysis revealed that the mean bias of the manual annotation and self-report compared to ZM was within $\pm$6 min with LOA no larger than a span of $\pm$45 min. Finally, the probability density distribution plots of the differences between the in-bed estimates of the manual raters and self-report compared to ZM were comparable in terms of the symmetry, spread around zero, and positioning of outliers.

The excellent performance of the prospective sleep diaries in the current study may be explained by the synchronized use of ZM and the sleep diaries. Thus, having the subject manually initiate and end the ZM recording every morning and night will make it easier for the subject to recall the time going to bed and time out of bed, thus, avoiding much of the usual discrepancy between objectively and subjectively measured sleep duration [28]. We would not expect to see such good agreement between the manual annotation and sleep diaries as well as ZM and sleep diaries if the participants wre instructed to exclusively log sleep using subjective measurements without protocol disturbances as anchor time points.

When compared to ZM, we found that the manual annotation of the in-bed period deviated more when estimating the going to bed timestamp. This could be caused by the difficulty the raters had in discriminating between inactive behaviors before bedtime and actual time in bed. However, these discrepancies may be minimized by further training of the rater's ability to distinguish between inactive behaviors; however, this poses the most important limitation to the manual annotation method. Nevertheless, the accuracy obtained in the present study is reassuring as it is achieved based on little preliminary formal training or briefing of the raters involved. In that sense, most of the work when determining the in-bed period is self-explanatory when provided with the information, given by the signal features selected in the present study in Audacity. This is further supported by the excellent ICC between the three manual raters. However, there appears to be a slight learning curve as the LOAs are consistently narrower during round 2 of the manual scoring compared to round 1. This is also evident in the density plots, which display a greater spread around zero during round 1 compared to round 2. This suggests that more than two rounds of manual scoring may homogenize the results further or that the raters may benefit from revisiting the annotations from the beginning of the first round of annotations. Alternatively, a form of training may be profitable before the actual annotation takes place. Further research is warranted to investigate methods to optimize the homogenization of the manual annotations. Nevertheless, the evidence suggests that supervised machine learning, given a large amount of labeled data, is resistant to label noise [29], which means that the tradeoff for accuracy in favor of the sheer volume of labeled data may be preferable. This further advocates for the use of the manual annotation methodology in data sets with no self-reported sleep or other measures of interest without labels.

There are other labeling tools for annotating time series data (e.g., Label Studio (Heartex Organization, San Francisco, CA, USA) [30] or Visplore (Visplore GmbH, Vienna, Austria) [31]); however, we found that Audacity was well suited for this specific task. Label Studio, for instance, may have difficulties handling week-long accelerometer data with 100 M+ entries, and Audacity is perfectly suited with its ability to seamlessly handle and navigate very large data structures. Furthermore, our feature selection was based on domain knowledge with the purpose of providing the right combination of features in limited number to avoid overflowing the raters with redundant information. This methodology can be extended to other behaviors, e.g., walking, which would likely require a different set of features. Although we do not provide clear-cut guidelines for the process of annotating the data, the raters in the present study were able to gain the right insights. Furthermore, the labeling of data is a step that inherently requires common knowledge in human behavior, and if the labels can sufficiently be described based on formal rules,

one can question whether the training of an AI model is necessary at all. Nevertheless, we suggest that further research investigating which features are the most important for successful annotations and, likewise, examine the effect of other sets of features that might provide important knowledge that could further facilitate the use of manual annotation of accelerometry time series data.

To date, most studies that have investigated the validity of actigraphy and self-report compared to PSG or EEG-based methodologies, have routinely evaluated sleep parameters, such as the total sleep time, wake after sleep onset, sleep latency, and sleep efficiency. These measures are often an aggregate measure that includes sleep onset and wake onset, which would be comparable to the to-bed and out-of-bed timestamps in the current study. Moreover, the precision of these time points is not evaluated and, thus, is difficult to compare to the measurements of the current study. The novel methodology for annotating the time in bed in the present study, however, provides ICC values on par with or better than previous studies comparing actigraphy sleep parameters to PSG [7,32]. Furthermore, one study presented mean absolute errors of 39.9 min and 29.9 min for the sleep onset time and waking up time, respectively, and 95% limits of agreement above $\pm$3 h for sleep duration when comparing an algorithm to PSG [10]. Furthermore, the ability to precisely estimate the timestamp of certain events compared to durations of specific behaviors leaves less room for error in the effort to obtain good agreements. Additionally, our manual methodology performs strongly across age and gender as the included subjects in the present study consist of both children and adults of both genders. This suggests that our manual annotation method is accurate irrespective of the inclusion of different developmental age groups and genders and their specific behavior and that it may be a more precise tool for estimating exact time points compared to present state automated methodologies. Traditionally, the accuracy of the assessment of sleep parameters is highly dependent on the target population, and thus we view the current results with plausible high generalizability to populations of normal sleepers.

Identifying periods of sleep (rather than simply lying) is an important component of a 24 h behavior profile, and many studies examining sleep detection based on conventional accelerometers involve asking the participant to record their time in bed, sleep onset, and waking up time [33–35]. The use of self-reported measures of sleep may be replaced by annotating in-bed manually, thereby, lowering the participant burden and avoiding the inevitable recall bias associated with self-reported measures. Therefore, an important application of the manual annotation methodology using Audacity is that it can, with no difficulty, be employed on free-living data. Moreover, the application of our proposed methodology is manifold. Other suitable use cases could be the annotation of non-wear time, manual clock synchronization of several different devices, examining the validity of raw data, and more. Additionally, it is not limited to actigraphy data but can be utilized on a wide variety of multi-channel data for an increased overview, including orientation (gyroscopic data), temperature, battery voltage, and light as examples. Finally, Audacity provides a fluid workflow even with very large multi-channel data files with the ability to swiftly zoom to every resolution needed and scroll through time with no lag and add labels that advocate the use of Audacity as a standard tool for researchers working with raw data and machine learning. For these purposes, the implementation of the Audacity-methodology on raw accelerometer data may help drive the development of future human behavior research.

Although the work of changing from raw sensor data to operational predictive models using labelled data has been the standard method for years, no previous studies have proposed a methodology that enables researchers to make optimal use of their available accelerometry data. We show that with a few well-selected features, the annotation of sleep is comparable to EEG-based sleep classification hardware. However, it is important to note that we do not currently recommend that our proposed methodology of manually annotating in-bed time is to be used as a replacement of other more well-established techniques for estimating sleep, e.g., EEG- or tracheal-sound-based options in ongoing

studies. Though, it may serve as a post-hoc procedure to enrich already collected data with a measure of sleep.

The strengths of this study include the continuous data collection of both accelerometry, sleep diary, and ZM in the home of the participants during multiple consecutive days of recordings making it high-quality free-living data. The limitations include the limited rater generalizability as the three manual raters were fixed and not randomly chosen from a larger population of eligible raters to accommodate different characteristics. However, due to the scarce pre-briefing instructions on how to label the raw data, we suggest that this methodology is generalizable to at least other researchers working with accelerometer data.

A natural next step would be to develop and validate a procedure similar to what is available with sleep annotation with EEG (AASM Scoring Manual [36]) and propose guidelines to make the manual annotation methodology accessible to persons with limited experience within the field of accelerometry. To record true free-living behavior using participant-mounted devices is difficult and wearing the ZM during sleep may affect the behavior of the participants and, thus, poses as a limitation of the study.

Additionally, although the criterion measure in the current study is validated against PSG, it would have been more optimal to use PSG as a criterion measure. Finally, we did not incorporate napping behavior in the current study as we focused on sleep in relation to circadian rhythms. Further research is needed to validate the manual methodology for use in detecting napping behavior.

5. Conclusions

In conclusion, our results show that the manual annotation of the in-bed period from thigh- and hip-worn accelerometer data using Audacity demonstrated good agreement with a minimal mean bias and acceptable limits of agreement for the time to bed and out of bed when compared to the same estimates assessed with the use of an objective EEG-based sleep device and prospective sleep dairies. Furthermore, the manual annotation was highly reliable with excellent inter- and intra-rater agreement and has an accuracy with the EEG-based assessment similar to the sleep diaries.

The study shows that the manual annotation can be used on already collected raw data when not accompanied by sleep records. This will facilitate the additional use of free-living data resources and, thus, could increase the amount of available training data when employing data-demanding machine learning algorithms. This has the potential to improve the generalizability of these algorithms in assessing human behavior from objective recordings.

Author Contributions: Conceptualization, N.C.M., E.L.S. and J.C.B.; methodology, N.C.M., E.L.S. and J.C.B.; validation, J.P., J.C.B. and E.L.S.; formal analysis, E.L.S.; investigation, E.L.S., J.P. and J.C.B.; data curation, A.G. and J.P.; writing—original draft preparation, E.L.S.; writing—review and editing, J.C.B., J.P., N.C.M. and A.G.; visualization, E.L.S.; supervision, J.C.B.; funding acquisition, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: TrygFonden (grant number 130081) and European Research Council (grant number 716657).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available due to the general data protection regulations but will be shared on reasonable request using a safe platform by the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fiorillo, L.; Puiatti, A.; Papandrea, M.; Ratti, P.-L.; Favaro, P.; Roth, C.; Bargiotas, P.; Bassetti, C.L.; Faraci, F.D. Automated sleep scoring: A review of the latest approaches. *Sleep Med. Rev.* **2019**, *48*, 101204. [CrossRef] [PubMed]

2. Van der Ploeg, T.; Austin, P.C.; Steyerberg, E.W. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* **2014**, *14*, 137. [CrossRef] [PubMed]
3. Chaput, J.-P.; Gray, C.E.; Poitras, V.J.; Carson, V.; Gruber, R.; Birken, C.S.; MacLean, J.E.; Aubert, S.; Sampson, M.; Tremblay, M.S. Systematic review of the relationships between sleep duration and health indicators in the early years (0–4 years). *BMC Public Health* **2017**, *17*, 855. [CrossRef] [PubMed]
4. Chaput, J.-P.; Gray, C.E.; Poitras, V.J.; Carson, V.; Gruber, R.; Olds, T.; Weiss, S.K.; Connor Gorber, S.; Kho, M.E.; Sampson, M.; et al. Systematic review of the relationships between sleep duration and health indicators in school-aged children and youth. *Appl. Physiol. Nutr. Metab.* **2016**, *41*, S266–S282. [CrossRef] [PubMed]
5. St-Onge, M.-P.; Grandner, M.A.; Brown, D.; Conroy, M.B.; Jean-Louis, G.; Coons, M.; Bhatt, D.L. Impact on Lifestyle Behaviors and Cardiometabolic Health: A Scientific Statement From the American Heart Association. *Circulation* **2016**, *134*, e367–e386. [CrossRef] [PubMed]
6. Gruber, R.; Carrey, N.; Weiss, S.K.; Frappier, J.Y.; Rourke, L.; Brouillette, R.T.; Wise, M.S. Position Statement on Pediatric Sleep for Psychiatrists. *J. Can. Acad. Child Adolesc. Psychiatry* **2014**, *23*, 174–195.
7. Haghayegh, S.; Khoshnevis, S.; Smolensky, M.H.; Diller, K.R. Application of deep learning to improve sleep scoring of wrist actigraphy. *Sleep Med.* **2020**, *74*, 235–241. [CrossRef] [PubMed]
8. Vaughn, B.V.; Giallanza, P. Technical review of polysomnography. *Chest* **2008**, *134*, 1310–1319. [CrossRef] [PubMed]
9. Sundararajan, K.; Georgievska, S.; te Lindert, B.H.W.; Gehrman, P.R.; Ramautar, J.; Mazzotti, D.R.; Sabia, S.; Weedon, M.N.; van Someren, E.J.W.; Ridder, L.; et al. Sleep classification from wrist-worn accelerometer data using random forests. *Sci. Rep.* **2021**, *11*, 24. [CrossRef] [PubMed]
10. Van Hees, V.T.; Sabia, S.; Jones, S.E.; Wood, A.R.; Anderson, K.N.; Kivimäki, M.; Frayling, T.M.; Pack, A.I.; Bucan, M.; Trenell, M.I.; et al. Estimating sleep parameters using an accelerometer without sleep diary. *Sci. Rep.* **2018**, *8*, 12975. [CrossRef] [PubMed]
11. Van de Water, A.T.M.; Holmes, A.; Hurley, D.A. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography—A systematic review. *J. Sleep Res.* **2011**, *20*, 183–200. [CrossRef] [PubMed]
12. Van Hees, V.T.; Sabia, S.; Anderson, K.N.; Denton, S.J.; Oliver, J.; Catt, M.; Abell, J.G.; Kivimäki, M.; Trenell, M.I.; Singh-Manoux, A. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. *PLoS ONE* **2015**, *10*, e0142533. [CrossRef]
13. Madsen, M.T.; Rosenberg, J.; Gögenur, I. Actigraphy for measurement of sleep and sleep-wake rhythms in relation to surgery. *J. Clin. Sleep Med.* **2013**, *9*, 387–394. [CrossRef] [PubMed]
14. Schwab, K.E.; Ronish, B.; Needham, D.M.; To, A.Q.; Martin, J.L.; Kamdar, B.B. Actigraphy to Evaluate Sleep in the Intensive Care Unit. A Systematic Review. *Ann. Am. Thorac. Soc.* **2018**, *15*, 1075–1082. [CrossRef] [PubMed]
15. Barouni, A.; Ottenbacher, J.; Schneider, J.; Feige, B.; Riemann, D.; Herlan, A.; Hardouz, D.E.; McLennan, D. Ambulatory sleep scoring using accelerometers—distinguishing between nonwear and sleep/wake states. *PeerJ* **2020**, *8*, e8284. [CrossRef] [PubMed]
16. Younes, M.; Raneri, J.; Hanly, P. Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability. *J. Clin. Sleep Med.* **2016**, *12*, 885–894. [CrossRef] [PubMed]
17. Dafna, E.; Tarasiuk, A.; Zigel, Y. Sleep-Wake Evaluation from Whole-Night Non-Contact Audio Recordings of Breathing Sounds. *PLoS ONE* **2015**, *10*, e0117382. [CrossRef]
18. Montazeri Ghahjaverestan, N.; Akbarian, S.; Hafezi, M.; Saha, S.; Zhu, K.; Gavrilovic, B.; Taati, B.; Yadollahi, A. Sleep/Wakefulness Detection Using Tracheal Sounds and Movements. *Nat. Sci. Sleep* **2020**, *12*, 1009–1021. [CrossRef] [PubMed]
19. Rasmusen, M.; Pedersen, J.; Olesen, L.; Kristensen, P.; Brønd, J.; Grøntved, A. Feasibility of two screen media reduction interventions: Results from the SCREENS pilot trial. *PLoS ONE* **2021**, *16*, e0259657.
20. Rasmussen, M.G.B.; Pedersen, J.; Olesen, L.G.; Brage, S.; Klakk, H.; Kristensen, P.L.; Brønd, J.C.; Grøntved, A. Short-term efficacy of reducing screen media use on physical activity, sleep, and physiological stress in families with children aged 4–14: Study protocol for the SCREENS randomized controlled trial. *BMC Public Health* **2020**, *20*, 380. [CrossRef] [PubMed]
21. Pedersen, J.; Rasmussen, M.G.B.; Olesen, L.G.; Kristensen, P.L.; Grøntved, A. Self-administered electroencephalography-based sleep assessment: Compliance and perceived feasibility in children and adults. *Sleep Sci. Pract.* **2021**, *5*, 8. [CrossRef]
22. Jaeschke, L.; Steinbrecher, A.; Jeran, S.; Konigorski, S.; Pischon, T. Variability and reliability study of overall physical activity and activity intensity levels using 24 h-accelerometry-assessed data. *BMC Public Health* **2018**, *18*, 530. [CrossRef]
23. Wang, Y.; Loparo, K.A.; Kelly, M.R.; Kaplan, R.F. Evaluation of an automated single-channel sleep staging algorithm. *Nat. Sci. Sleep* **2015**, *7*, 101–111. [CrossRef]
24. Kaplan, R.F.; Wang, Y.; Loparo, K.A.; Kelly, M.R.; Bootzin, R.R. Performance evaluation of an automated single-channel sleep-wake detection algorithm. *Nat. Sci. Sleep* **2014**, *6*, 113–122. [CrossRef] [PubMed]
25. Skotte, J.; Korshøj, M.; Kristiansen, J.; Hanisch, C.; Holtermann, A. Detection of Physical Activity Types Using Triaxial Accelerometers. *J. Phys. Act. Health* **2014**, *11*, 76–84. [CrossRef]
26. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
27. Bland, J.M.; Altman, D.G. Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* **1999**, *8*, 135–160. [CrossRef]
28. Aili, K.; Åström-Paulsson, S.; Stoetzer, U.; Svartengren, M.; Hillert, L. Reliability of Actigraphy and Subjective Sleep Measurements in Adults: The Design of Sleep Assessments. *J. Clin. Sleep Med.* **2017**, *13*, 39–47. [CrossRef] [PubMed]

29. Rolnick, D.; Veit, A.; Belongie, S.; Shavit, N. Deep Learning is Robust to Massive Label Noise. *arXiv* **2018**, arXiv:170510694.
30. Label Studio–Open Source Data Labeling. Available online: https://labelstud.io (accessed on 6 December 2021).
31. Visplore–Software for Visual Time Series Analysis. Available online: https://visplore.com/home (accessed on 6 December 2021).
32. Yavuz-Kodat, E.; Reynaud, E.; Geoffray, M.-M.; Limousin, N.; Franco, P.; Bourgin, P.; Schroder, C.M. Validity of Actigraphy Compared to Polysomnography for Sleep Assessment in Children With Autism Spectrum Disorder. *Front. Psychiatry* **2019**, *10*, 551. [CrossRef] [PubMed]
33. Littner, M.; Kushida, C.A.; Anderson, W.M.; Bailey, D.; Berry, R.B.; Davila, D.G.; Hirshkowitz, M.; Kapen, S.; Kramer, M.; Loube, D.; et al. Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: An update for 2002. *Sleep* **2003**, *26*, 337–341. [CrossRef] [PubMed]
34. Lockley, S.W.; Skene, D.J.; Arendt, J. Comparison between subjective and actigraphic measurement of sleep and sleep rhythms. *J. Sleep Res.* **1999**, *8*, 175–183. [CrossRef]
35. Girschik, J.; Fritschi, L.; Heyworth, J.; Waters, F. Validation of self-reported sleep against actigraphy. *J. Epidemiol.* **2012**, *22*, 462–468. [CrossRef] [PubMed]
36. AASM Scoring Manual-American Academy of Sleep Medicine. Available online: https://aasm.org/clinical-resources/scoring-manual (accessed on 8 November 2021).

Appendix II

Generalizability and Performance of Methods to Detect Non-Wear With Free-Living Accelerometer Recordings

DOI: https://doi.org/10.1038/s41598-023-29666-x

scientific reports

Check for updates

OPEN

Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings

Esben Lykke Skovgaard✉, Malthe Andreas Roswall, Natascha Holbæk Pedersen, Kristian Traberg Larsen, Anders Grøntved & Jan Christian Brønd

Wearable physical activity sensors are widely used in research and practice as they provide objective measures of human behavior at a low cost. An important challenge for accurate assessment of physical activity behavior in free-living is the detection non-wear. Traditionally, heuristic algorithms that rely on specific interval lengths have been employed to detect non-wear time; however, machine learned models are emerging. We explore the potential of detecting non-wear using decision trees that combine raw acceleration and skin temperature, and we investigate the generalizability of our models, traditional heuristic algorithms, and recently developed machine learned models by external validation. The Decision tree models were trained using one week of data from thigh- and hip-worn accelerometers from 64 children. External validation was performed using data from wrist-worn accelerometers of 42 adolescents. For non-wear episodes longer than 60 min, the heuristic algorithms performed the best with F1-scores above 0.96. However, regarding episodes shorter than 60 min, the best performing method was the decision tree model including the six most important predictors with F1 scores above 0.74 for all sensor locations. We conclude that for classifying non-wear time, researchers should carefully select an appropriate method and we encourage the use of external validation when reporting on machine learned non-wear models.

Over the past few decades body-worn motion sensors have been used to study human physical activity behavior as these devices have been shown to provide a robust method for measuring the characteristics of free-living movement[1–3]. The use of activity monitors based on accelerometry offers objective data capture and the ability to assess intensity of movement with minimal participant burden and at high cost efficiency[4]. Different methods are used to attach the devices on the subject and many protocols allow the subject to detach the device for water activities, sleep, or sports for which wearing the device can cause an injury. Non-wear time is defined as periods of not wearing the devices which can have important consequences for the outcomes derived from the acceleration measurements. Since non-wear time is equivalent to missing data, researchers may exclude the non-wear time from their analysis or perform imputation of the missing data using various methods such as zero-inflated Poisson and Log-normal distributions[5]. However, imputing non-wear time can introduce bias, especially for longer periods of non-wear time, because every imputation method is based on assumptions about the data, which may not be accurate. Therefore, the optimal classification and handling of these non-wear periods is important for providing researchers with high quality estimates of the subject's physical activity behavior during free-living.

The classification of non-wear periods in the physical activity measurements can be obtained by having the subjects keep an individual log diary although this method is cumbersome for the subject and is potentially error prone[6]. To reduce the burden on the subject and increase accuracy, researchers have employed different rule-based methods and more advanced algorithms to classify non-wear time. The earliest rule-based methods were developed for ActiGraph counts data which classify non-wear time as the periods of data with consecutive zero counts exceeding a specified duration[7–9]. Although the simplicity of these methods is considered a strength, it has been shown that alternating the length of the time interval of consecutive zero counts result in differences in physical activity and sedentary behavior aggregates of up to 10%[10]. Moreover, Until recently, the algorithms used to convert acceleration data into counts were proprietary, which would impede the transparency of the research

Research Unit for Exercise Epidemiology, Department of Sports Science and Clinical Biomechanics, Centre of Research in Childhood Health, University of Southern Denmark, 5230 Odense, Denmark. ✉email: eskovgaard@health.sdu.dk

field. Finally, wear-time inclusion criteria have been shown to vary depending on non-wear settings with age and obesity level eliciting relative differences in non-wear time[11]. This makes the consecutive zeros algorithms sub-optimal and hinders comparisons across studies with differing populations and non-wear settings[11].

The technological advances during the last couple of decades of accelerometers have provided researchers with the ability to store raw accelerations (in gravity units), which increases the granularity of the data and potentially the ability to accurately classify non-wear periods. Different algorithms have been developed to detect non-wear time in raw accelerometer data which typically utilize a standard deviation threshold but also in combination with surface skin temperature[14–16]. These heuristic approaches have proven to be generalizable across different populations, device brands and wear-sites, but having access to data of increasing quantity and quality does in principle not improve their performance as is possible with a machine learning approach which can improve as data becomes available. Furthermore, the use of simple duration-based algorithms carries a risk of falsely misclassifying true non-wear as inactivity as they are restricted by time length-specific intervals and any non-wear episodes shorter than the interval cannot be detected. In the current study, we define *generalizability* as the extent to which the performance of a model can be applied to out-of-sample unseen data, i.e., data beyond the specific sample used in the development of the model. If a model has low generalizability, its findings may be limited in their applicability to other populations or settings.

Recently, studies have investigated the potential of classifying non-wear using raw accelerometer data in conjunction with machine learning like random forests[17] and deep learning techniques[18]. The purpose of using machine learning algorithms is to learn patterns from the training data and to approximate the complex model which best describes the relationship between the included predictors and the outcome. In this process, the trade-off between model variance and bias must be taken into consideration. Variance refers to the amount by which the approximated function would change if it were estimated on a different training dataset and bias refers to the error that is introduced by approximating a complicated problem with a model that assumes a simple relationship (e.g., linear regression estimating a highly non-linear function). Thus, high bias results in underfitting of the model and high variance results in overfitting. In general, highly flexible, and complex models with high variance/low bias pay a lot of attention to the training data and fail to generalize to unseen data. As a result, such models perform very well on the training data but have high prediction error on unseen data, i.e., the model is overfitted to the training data. Contrary, simpler models with less flexibility are prone to underfit the training data due to high bias and low variance. Therefore, the balance between overfitting and underfitting is of foremost importance when the predictive performance of a machine learned model is to be used on an unseen data source. Although the models utilizing complex machine learning algorithms have shown to perform exceptional on testing data, it is unknown to which degree the models perform on external unseen data. Although this can be viewed as a shortcoming of studies reporting performance metrics with no external validation, it is important to recognize that the performance of a given machine learning or deep learning model will always, in principle, be unknown on out-of-distribution data sources. However, as most non-wear detection methods are developed with the goal in mind of being device, placement, and population agnostic, external validation will be of value as an indication of model generalization. Despite this, there exists several reasons to why researchers are not employing external validation, including lack of out-of-distribution data sources and/or the desire to incorporate all available data into the training of the models to capture the most information. A few studies have been utilizing surface skin temperature in combination with raw acceleration for the classification of non-wear episodes[14,16]. However, the performance and generalizability of adding the surface skin temperature for the classification of non-wear with advanced machine learning methods has not been investigated.

To date, accurate classification of non-wear time in raw accelerometer data still has potential for improvements despite advancements in sensor technology and related software. This begs the question, as a tool for the essential first step when analyzing accelerometer data, i.e., classifying non-wear time, what heuristic algorithm or machine learned model will perform the best on unseen data?. To answer this, we created three datasets of raw accelerometer data with correctly labeled wear- and non-wear time as ground truth including surface skin temperature measurements. In specific, we aimed to (1) train three decision tree models on accelerometer data from thigh and hip-worn accelerometers for the classification of non-wear time in raw accelerometer data and evaluate the importance of surface skin temperature and minimizing the number of predictors provided to the model and (2) evaluate the performance of machine-learned models and simple heuristic algorithms across datasets of varying age ranges for the classification of non-wear time in raw accelerometer data.

Background

We included in total four additional non-wear classification methods to evaluate generalizability and to compare the performance with the three new algorithms developed. These existing methods are carefully selected on the premise that we wished to examine a spectrum of method flexibility such that the simplest (and most widely used) and the most recent and complex techniques are included.

Consecutive zeros-algorithm (*cz_60*). A variety of consecutive zero-algorithms for count-based accelerometer data have been developed over the years for detecting periods of non-wear within specified time intervals, such as 30-, 60-, or 90-min intervals[7,9,19]. Furthermore, non-wear algorithms using raw acceleration have been developed by van Hees and colleagues with a 30 min interval[12] who later extended their work with a 60-min interval algorith[15], and one with a 135-min interval and tuned hyperparameters by Syed et al.[20] Here we employ a simple implementation of this concept of detecting no movement based on Actigraphy counts with an algorithm that detects only zero counts for a minimum of 60 consecutive minutes. Actigraphy counts are generated with a deadband of 68 mg, making it the minimum detectable acceleration threshold.

Heuristic algorithm (*heu_alg*). This algorithm is described in detail by Rasmussen and colleges[15] and utilizes a combination of raw acceleration and surface skin temperature. Periods longer than 120 min with accelerations below 20 mg are always identified as non-wear time and periods between 45 and 120 min are identified as non-wear if the temperature is below an individually estimated non-moving temperature threshold. Finally, the algorithm detects non-wear periods of 10–45 min in duration only if the end of the non-wear period is within the expected awake time.

Random forests model (*sunda_RF*). The non-wear classification method described by Sundararajan et al. is based on a random forest ensemble model which was trained on raw accelerometer data from 134 subjects aged 20–70 years. The subjects wore an accelerometer on their wrist during a single overnight PSG recording. The ground truth labels for non-wear time assumed that the accelerometer was worn only during the PSG recording. Only in epochs where the standard deviation in the acceleration signal per 15 min was larger than 13.0 mg outside the PSG recording was labelled as wear time. As input data, 36 predictors were constructed, and a nested cross-validation approach was employed to obtain generalization performance and to tune hyperparameters.

Deep convolutional neural network (*syed_CNN*). The non-wear classification method described by Syed et al.[18] is based on a deep convolutional neural network using a novel approach which is different from the other methods included. Initially, all candidate non-wear episodes are identified using a standard deviation threshold. Then, rather than inspecting acceleration within the candidate non-wear intervals as the other methods do, this approach examined the signal shape of the raw acceleration right before and right after an episode of non-wear time using a convolutional neural network (CNN). In essence, they developed a CNN that can infer non-wear time by detecting when the accelerometer is detached and when it is mounted back on again. In this study, we used a window length on each side of the candidate non-wear period of 10 s, as this produced the best results. The training data for developing the CNN was collected from hip-worn accelerometers of 583 participants aged 40–84 years (mean = 62.74; SD = 10.25).

Methods

Investigating the classification performance and generalizability of the non-wear classification methods was done by evaluating the performance with data collected in free-living using accelerometers placed at the wrist, thigh, and hip. The data was established by combining data collected in the Physical Activity in Schools After the Reform study (PHASAR)[21] which provides data for both hip- and thigh, and an in-house data validation study which collected acceleration with wrist-worn accelerometers. An outline of how the data was used is presented in Fig. 1. Using the three data sets in this way, we ensured that the previously established non-wear classification methods were evaluated on an external data set. We also ensured that our own decision tree models were evaluated on an external independent dataset with data and wear location, which was not used for developing the decision tree models. Thus, all included machine learned models are evaluated with a test dataset collected from anatomical positions both included and not included in the development of the models.

Data sources. With both the PHASAR and in-house validation study, raw acceleration data in conjunction with surface skin temperature was recorded by the Axivity AX3 accelerometer (Axivity Ltd., Newcastle upon Tyne, UK). The Axivity AX3 provides a dynamic range of ± 8 g (1 g = 9.81 ms^{-2}), with physical dimensions 23 mm × 32.5 mm × 7.6 mm and weighing only 11 g. The Axivity AX3 device stores acceleration in gravity units (g) along three axes (vertical, mediolateral, and anteroposterior) with a selectable sampling frequency of 6.25–3200 Hz. The sampling frequency was set to 50 Hz with the PHASAR study and 25 Hz with the in-house validation study. All acceleration data was resampled into 30 Hz for both studies.

The PHASAR study is comprised of a population-based sample of more than 2000 school-aged children from 31 public schools in Denmark[21]. The objectively measured physical activity data from hip- and thigh-worn accelerometers were obtained in the PHASAR study during 2017–2018 and consisted of data from 1315 (49%) boys and 1,358 (51%) girls aged 8.1–17.9 years old (mean = 12.14, SD = 2.40). The accelerometers were worn at two anatomical locations; one fixated to the body in a pocket attached to a belt worn around the waist, where the sensor was placed on the right hip with the USB connector facing away from the right side of the body. A second belt was worn around the right thigh midway between the hip and the knee, where the accelerometer was placed in a pocket with the USB connector facing away from the body. The devices were worn for 1 week (seven consecutive days) which corresponds to the recommended number of days required to reliably estimate habitual physical activity[22]. The in-house validation study is comprised of accelerometer recordings from 42 individuals, 21 boys and 21 girls. This data was collected on youth athletes aged 14.5–16.4 years old (mean = 15.4, SD = 0.37 years) in the Region of Southern Denmark, which were all enrolled in a tailored talent program at two public schools near the end of their high school period. The Axivity accelerometer was mounted to the non-dominant wrist using a rubber strap with an embedded socket for the sensor and was worn for 14 consecutive days. The data collection for the in-house validation study commenced in the early spring of 2021.

From the PHASAR cohort we included the raw accelerometer recordings of 64 randomly selected participants. Based on this dataset of 64 participants, a dataset with labeled episodes of true non-wear time was constructed by means of manual annotation, a methodology, which is described in detail elsewhere[23]. In short, the non-wear episodes were inferred by visually inspecting the raw accelerations in combination with skin temperature readings. From raw triaxial 30 Hz Axivity acceleration data, episodes of true non-wear with start and stop time stamps were manually labelled in each of the three constructed datasets and served as labels of ground truth in

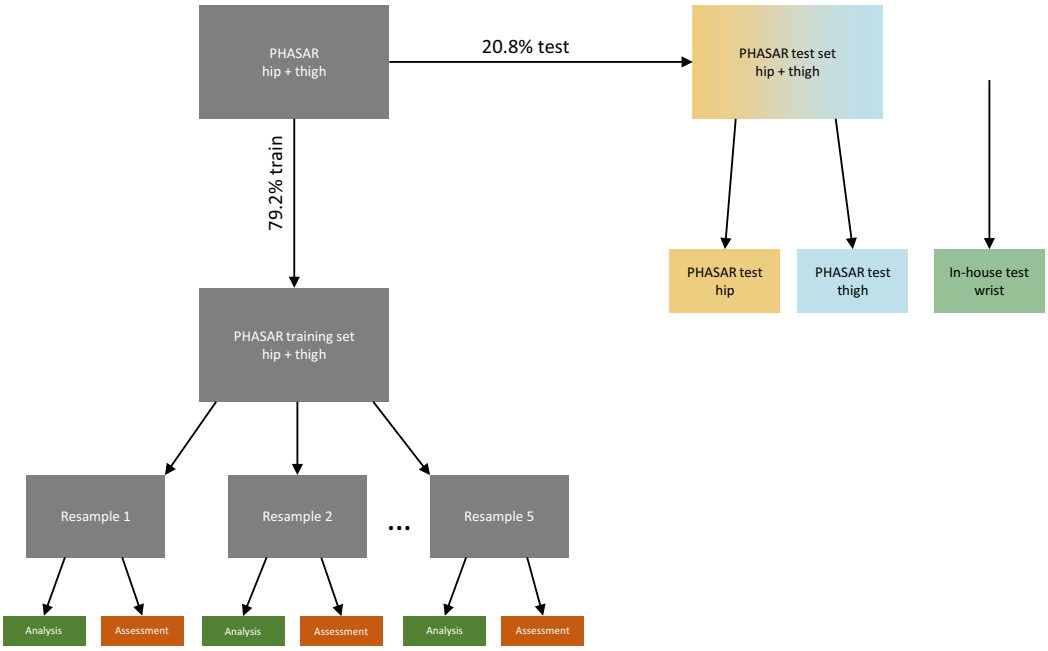

Figure 1. Flowchart of the splitting of the PHASAR dataset into train and test. The boxes on the left represent 79.2% of the PHASAR data for training in the five-fold resamples. The yellow and blue boxes on the right represent 20.2% of the PHASAR data for testing being split up into hip and thigh data while the green box is our in-house test dataset collected from wrist-worn devices.

later analyses. A second dataset was constructed in the exact same manner from the in-house validation study including all 42 youth athletes.

The PHASAR study was assessed by the Regional Committee on Health Research Ethics for Southern Denmark (ID: S-20170031) and deemed not eligible to undergo ethics review (documentation available on request to corresponding author) By Danish law, only research projects of biomedical character or projects that involve risks for participants need to have their ethics reviewed by a Regional Ethics Board. All other research projects can not apply for formal ethical approval. The in-house validation study was by the Research & Innovation Organization and legal department of University of Southern Denmark waived for ethical approval through the regional ethical committee. Both studies were carried out in accordance with the Danish Data Protection Agency (2015-57-0008) and all included participants, and/or their legal guardian(s) gave written informed consent. Furthermore, all methods were carried out in accordance with relevant guidelines and regulations (i.e., Declaration of Helsinki).

Development of decision tree models. For the development of our decision tree models, we extracted a total of 12 predictors from the raw accelerometer data from the PHASAR datasets including temperature, time of day, integer indicator variables for device placement and day of week and moving average statistics (see Table 1). Moving average predictors were aggregated in 10 s epochs.

The training of the decision tree models was performed on 79.2% of the PHASAR data stratified on non-wear and wear time including both the hip- and thigh-worn data (see Fig. 1. Furthermore, the data was split into training and test partitions such that participants in both partitions did not overlap which ensures that the algorithm did not learn any patterns specific to individual participant behavior. To optimize model hyperparameters and to minimize the risk of overfitting, a five-fold cross-validation scheme was used. The hyperparameters that were tuned were cost-complexity, tree depth and minimum number of data points required in a node for further splitting. For this, Latin hypercube sampling was used to construct a space-filling parameter grid with 10 levels that cover the parameter space such that any portion of the space had an observed combination that was not too far from it. The model with the most optimal hyperparameters was then trained on the full training dataset, still including both hip- and thigh-worn accelerometer data.

In the present work we present a full model (*tree_full*) including all predictors, a second model including the six best predictors (*tree_imp6*) based on permutation predictor importance (see Fig. 2), and a third model excluding the surface skin temperature (*tree_no_temp*) resulting in 5 times 10 models trained in the process per decision tree variant. The slight imbalance in the outcome variable of 55.8% wear time vs 44.2% non-wear time did not warrant the need to apply synthetic minority oversampling techniques (i.e., SMOTE) or other techniques to balance out the data.

Predictor	Description
Weekday	Day of week ([1:7])
time_day	Time of day (milliseconds)
Location	Device wear location: 0 = thigh, 1 = hip
macc_x	Moving average of the z axis acceleration
macc_y	Moving average of the y axis acceleration
macc_z	Moving average of the z axis acceleration
sdacc_x	Moving average of the standard deviation on the x axis acceleration
sdacc_y	Moving average of the standard deviation of the y axis acceleration
sdacc_z	Moving average of the standard deviation of the z axis acceleration
Sdmax	Maximum standard deviation
Incl	Inclination angle of the device in relation to the direction of the gravitational force
Temp	Surface skin temperature (degrees Celsius)

Table 1. Predictors derived from the raw sensor signals.

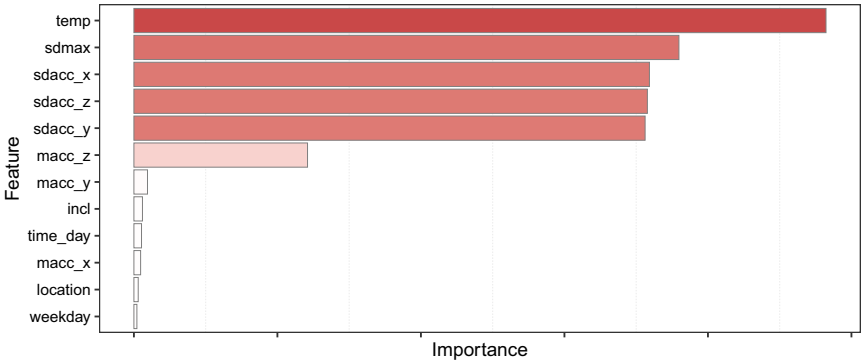

Figure 2. Predictor permutation importance plot for the decision tree including all predictors. The six most important predictors were used for training a second decision tree (*tree_imp6*), while all predictors excluding temperature were used to train a third decision tree (*tree_no_temp*).

Statistics. We calculated the classification performance to each ground truth test dataset comprising of more than 7 million 10 s epochs from 104 different subjects. True non-wear time inferred as non-wear time contributed to the true positives (TP), and true wear time inferred as wear time contributed to the true negatives (TN). Both TPs and TNs are required to obtain high accuracy of the non-wear time algorithm, as they are the correctly inferred classifications. True non-wear time inferred as wear time contributed to the false negatives (FN), and true wear time inferred as non-wear time contributed to the false positives (FP). The FPs, TPs, FNs, and TNs were calculated by looking at 10 s intervals of the acceleration data and comparing the inferred classification with the ground truth labels. From this we created a confusion matrix and derived overall accuracy as $\frac{TP+TN}{TP+TN+FP+FN}$, sensitivity as $\frac{TP}{TP+FN}$, precision as $\frac{TP}{TP+FP}$, and F1-score as $\frac{2TP}{2TP+FP+FN}$ to evaluate the classification performance of each non-wear detection technique. F1-score is the harmonic mean of precision and sensitivity with high F1-scores indicating good classification performance. Finally, we investigated the permutation predictor importance to understand the reason of better performances for each of the decision tree models.

For all analyses and model development, we used R version 4.1.2 (Bird Hippie) and RStudio version 2021.9.1.372 (Ghost Orchid) including the Tidymodels[24] suit of packages for machine learning tools, and the package rpart[25] as engine for the decision tree algorithm.

Results

In total across the three different wear locations, 1598 episodes of non-wear time were present in our gold standard datasets. The majority of these were non-wear episodes lasting ≥ 60 min, which accounted for 1148 (71.8%) episodes and had a mean duration of 794 min (SD = 1142), or approximately 13 h. Non-wear episodes lasting ≤ 60 min accounted for 450 (28.2%) of the episodes and had a mean duration of 26.4 min (SD = 16.4). Moreover, the shortest episodes (< 60 min) only constitute on average 1.3% of the total non-wear time across the three different wear-locations (see Table 2). Figure 3 shows the frequency distribution of episodes lasting < 60 min, and episodes lasting ≥ 60 min. The distribution of the short episodes of the PHASAR dataset was

Wear location	Mean[1]	Cumulated[1]	Proportion[2] (%)
< 60 min			
Hip	28	3202	1.13
Thigh	25	3975	1.40
Wrist	27	4691	1.32
≥ 60 min			
Hip	828	279,785	98.87
Thigh	776	280,294	98.60
Wrist	782	351,179	98.68

Table 2. Overview of non-wear episodes grouped in short and long non-wear episodes. [1]Aggregated in minutes. [2]Proportion of total non-wear time by wear location.

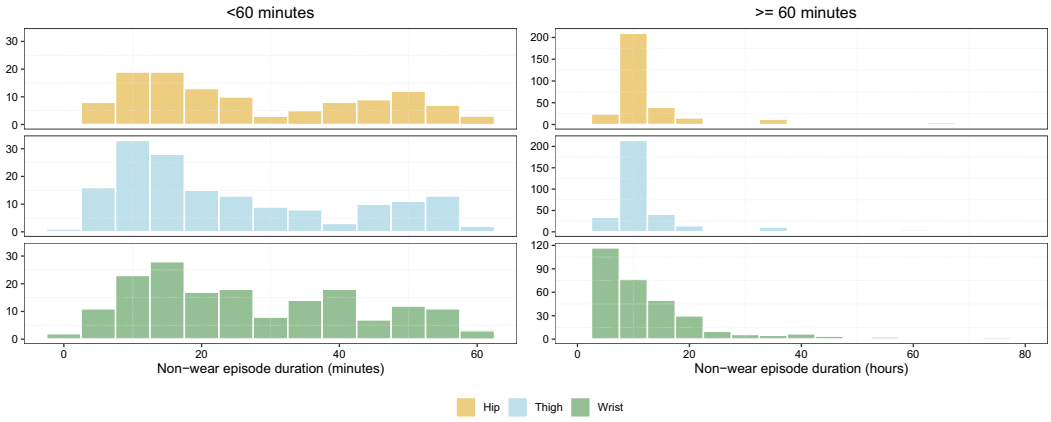

Figure 3. Distribution of the length of the non-wear episodes across hip, thigh, and wrist data. Distributions are shown for episodes shorter than 60 min and longer than 60 min.

bimodal while the episodes longer than 60 min were most frequent around 10 h. The short episodes of the in-house dataset from wrist-worn devices were uniform whereas the long episodes were highly right skewed.

Classification performance. Figure 4 shows an illustrative example of the outputs from the machine learned models and the rule-based algorithms shown comparatively with the ground truth non-wear time as shaded light blue background color. This particular case displays the general trend that the tree-based models are accurate but also volatile whereas the threshold-based methods (i.e., *Syed_CNN*, *heu_alg* and *cz_60*) are more stable. It is also evident in this recording, that the simple *cz_60* and *heu_alg* algorithms are not able to capture the short episodes.

Figure 5 summarizes all classification performance metrics for all methods included in this study. The CNN model by Syed et al. performed similar across all three datasets with an overall accuracy ranging from 75% to 80%. The CNN provides the highest sensitivity score with 93% to 96%. Furthermore, F1 scores ranging from 82% to 84% were obtained by the CNN which was impaired by a mediocre precision score. The random forests by Sundararajan et al. performed the best on the wrist data having an F1 score of 94% and accuracy of 93% but otherwise performed poorly on the hip and thigh data in comparison to the other methods with overall accuracy scores 56% and precision scores of 59% indicating the presence of many false positives (the misclassification of true wear time as non-wear time). The decision tree model without the surface skin temperature as a predictor was the worst performing of the decision tree models on the wrist data having an overall accuracy score of 72%. Although the model obtained an excellent sensitivity score of 98%, the poor precision of the model resulted in a lower F1 score of 81% compared to both the CNN and the random forests. The remaining two decision tree models, including the six most important predictors and the full model, scored excellent across all performance metrics and across all three datasets. Finally, the *heu_alg* and the *cz_60* algorithms scored near perfect across all metrics and datasets. Performance metrics on episodes less than or equal to 60 min in length are shown in Fig. 6. As expected, the results show that the simple consecutive zeros algorithm detects no non-wear (not depicted in Fig. 6). The deep learning model by Syed et al. performed poorly across all metrics only detecting 1–2% of all non-wear time resulting in F1 scores below 5%. The *heu_alg* algorithm elicited high precision scores but combined with poor sensitivity score the resulting F1 scores ranged from 12% to 16% across the different

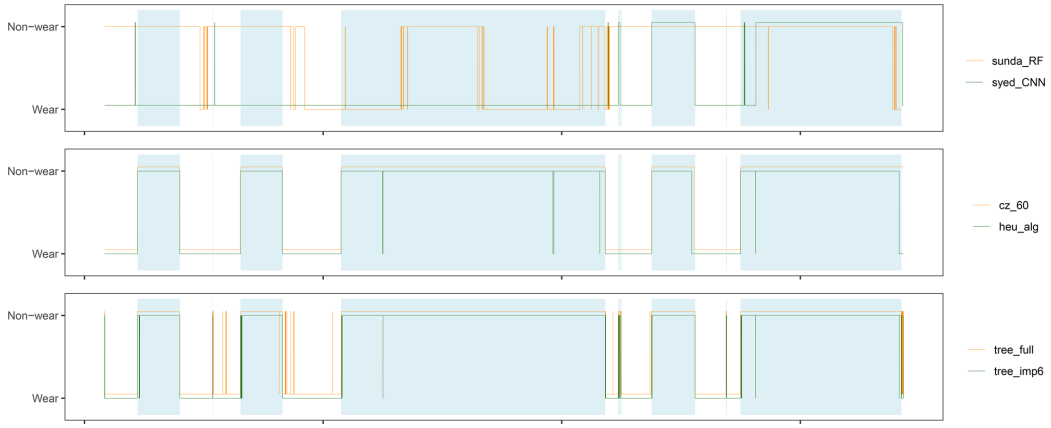

Figure 4. Visual example of the output of non-wear detection models and algorithms for a random person from the in-house wrist dataset (14 consecutive days). The light blue shade is ground-truth non-wear time. *Syed_CNN*, *cz_60*, and *tree_full* are vertically offset for easier interpretation.

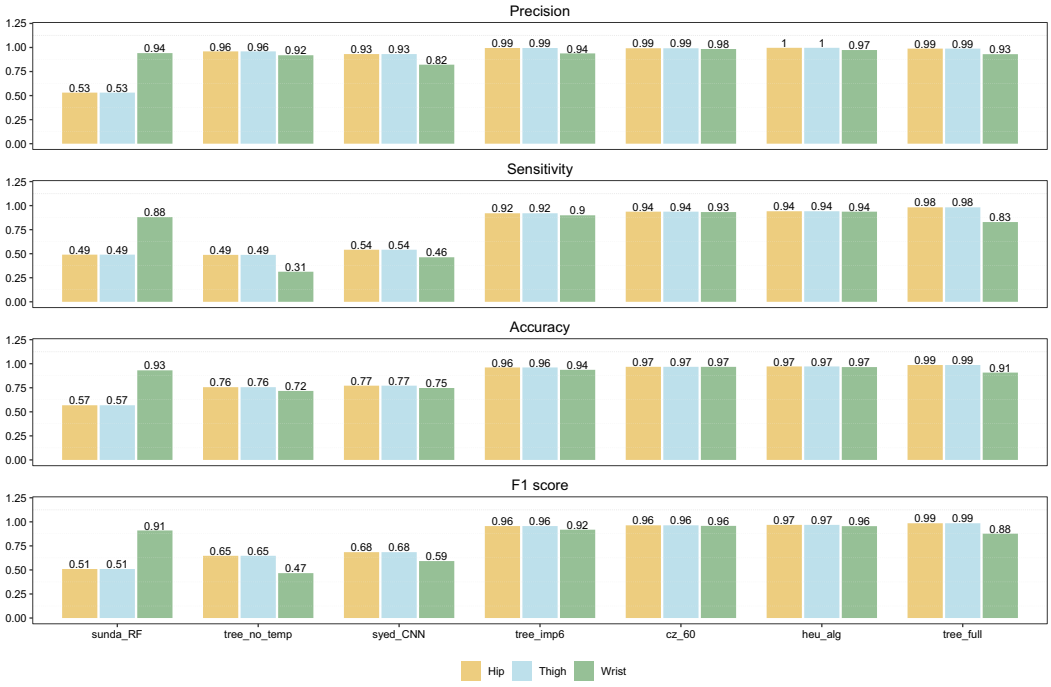

Figure 5. Classification performance metrics on all non-wear episodes for the seven included methods for classifying non-wear time. Metrics are shown for the three different ground-truth dataset including hip-worn, thigh-worn, and wrist-worn raw accelerometer data.

wear locations. The random forest model performed mediocre on thigh and wrist data, and poor on the hip data with F1 scores of 46%, 57% and 8%, respectively. out of the three decision tree models, the model including the six most important predictors performed the best with decent F1 scores ranging from 72% to 79%. The decision tree model trained on all predictors (*tree_full*) performed poorly on the hip data due to a low sensitivity score

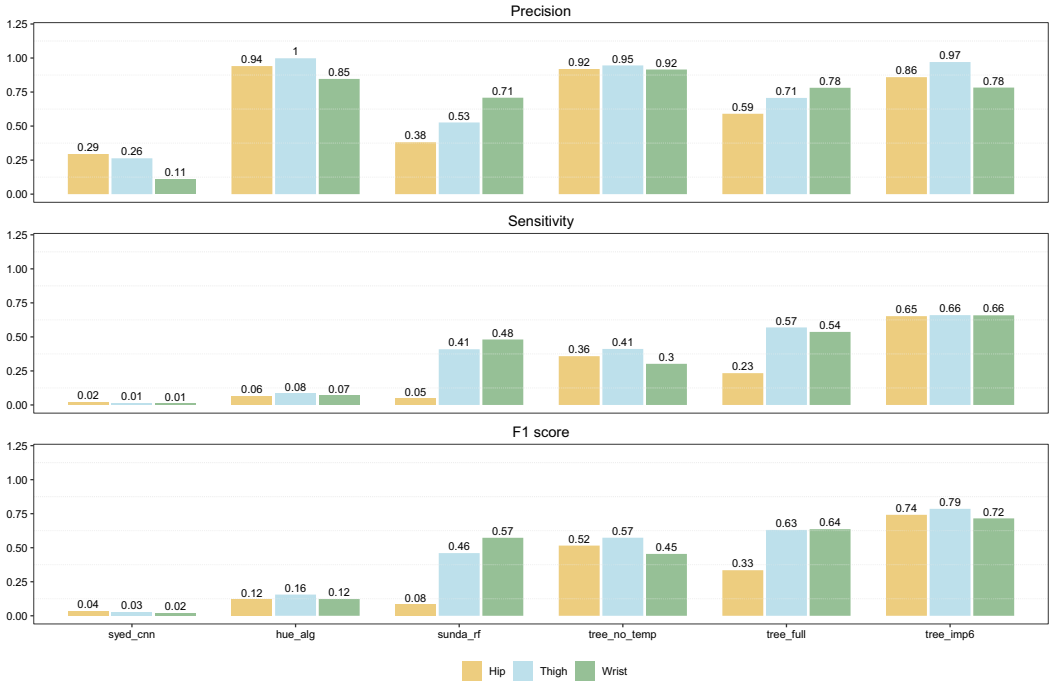

Figure 6. Classification performance for episodes no longer than 60 min in length. Metrics are shown for the three different gold-standard dataset including hip-worn, thigh-worn, and wrist-worn raw accelerometer data.

of 23%. The decision tree trained on all predictors, but surface skin temperature showed excellent precision but because of low sensitivity, the F1 scores ranged from 45% to 57%.

Discussion

Based on our results, the simplest methods (i.e., *cz_60* and *heu_alg*) for classifying non-wear episodes of lengths longer than 60 min performed excellent across all three wear locations closely followed by the decision tree models which included the surface skin temperature as a predictor. The random forest model also performed excellent on the wrist while the performance on the hip and thigh were mediocre. The deep learning and decision tree models without surface skin temperature as a predictor performed mediocre across all three sensor wear locations. As expected, when only examining the short non-wear episodes (< 60 min), the *cz_60* and the *heu_alg* algorithms were per default limited by their minimum episode duration of 60 and 20 min, respectively; thus, they performed poorly. The deep learning model performed poorly mainly due to a low sensitivity score resulting in many episodes falsely classified as non-wear. The random forest model performed poorly on the hip and mediocre on the thigh and wrist. We also observed mediocre performance for the decision tree model without temperature and the decision tree including all predictors. The best performing model on the short non-wear episodes was the decision tree trained on the six most important predictors. Lastly, inclusion of surface skin temperature to increase the predictive performance of non-wear time is supported by the results of this study.

The vast majority of the non-wear episodes in our ground truth datasets were longer than 60 min in length. In fact, the distribution of episode lengths revealed a peak around 10 h of duration which may be atypical compared to previous findings that show shorter episodes are the most numerous[10,22,26]. The characterization of the non-wear episodes within our data seems to favor the simple heuristic approaches for classifying non-wear time as the proportion of non-wear time lost to the limitations of employing a minimum window length are minuscule. Furthermore, the simple algorithms were able to obtain excellent precision scores indicating that no sedentary time or sleep was misclassified as being non-wear time which contrasts with several previous studies that have highlighted the difficulties performing this distinction[9,14,19,27–30]. Indeed, the present study included children and physically active adolescents which are known to spend less time sedentary and breaking up sedentary time more frequently[31,32]. This may accentuate the distinction between sedentary behavior and non-wear time suggesting that using a consecutive zeros algorithm may be considered best practice to capture non-wear episodes lasting longer than 60 min in children and adolescents across the wear locations included in this study, i.e., hip, thigh, and wrist. In addition, it is evident that the differences in physical activity behavior between children/adolescents and older adults will limit the ability to generalize models trained on specific age groups to other age groups.

However, this may be less applicable regarding the standardized procedure of mounting and unmounting the accelerometer, as is the case with the *syed_CNN* model.

The task of creating a model to classify non-wear time can be considered simple. The underlying decision boundary is likely close to linear; and thus, the more complex models included in the current study may end up over-fitting to the random variation that are present only in the specific dataset used for training which deteriorates the generalizability to unseen data. Therefore, we speculate that an adequately optimized logistic regression model would likely perform on par with the included methodologies as the separating linear hyperplane would be able to differentiate between wear and non-wear time. Thus, utilizing highly non-linear models for the classification of non-wear time may be unnecessarily complex if the goal is to create a machine learning model that is to be employed across various populations and wear sites. Alternatively, it is essential to include different wear locations and a variety of physical activity profiles in the training data.

To the best of our knowledge, including the surface skin temperature for the classification of non-wear time has only scarcely been investigated using machine learning techniques with a single study showing that acceleration in combination with rate-of-change in surface skin temperature can result in a robust decision tree model for the detection of non-wear time[33]. Previous studies have incorporated temperature in heuristic non-wear algorithms showing that predictive performance is to be gained[14,16] which is in accordance with our results that clearly demonstrate that including surface skin temperature improves the performance of the non-wear model. A critical aspect is the exact detection of the non-wear/wear onset and the slow temperature step response time of the sensor. Relying solely on temperature may introduce potential delays in the classification if the response time of the temperature sensor is slow whereas combining both temperature and acceleration data seems to be preferable[16]. We observed a 20 min step response of the Axivity temperature sensor, which might be explained by the way the devices casing has been designed. Additionally, the temperature is also dependent on the type of attachment method employed. When more material is positioned between the skin and the device, the longer the delay will be; thus, the machine learned models might ought to take type of sensor attachment into account. Moreover, device brands are likely to influence the temperature data collection as different temperature thresholds have been found to be optimal for distinct brands of devices, and as pointed out by Duncan et al. and Zhou et al. modifications were needed for the algorithms to work in different latitudes[14,16].

Processing accelerometry data, it would be preferable to use the same model across various wear locations and populations, hence; to further qualify the robustness of the generalization performance of the included methods in the present study, we incorporated a dataset from wrist worn devices. By introducing a dataset from wrist-worn devices, we ensure the performance metrics of our developed decision tree models are not inflated due to overfitting because of lack of variance between training and testing data. This procedure is known as *external validation* and involves using independently derived datasets to validate the performance of a model trained on initial input data. Due to the test dataset coming from an independent source, any predictor set that may have been falsely selected due to characteristics of the input training data (e.g., technical or sampling bias) would likely fail. Hence, a positive performance in external validation is regarded as a proof of generalizability[34]. The logic behind external validation is sound: data taken from separate sources have less in common, but nonetheless may capture useful domain-relevant aspects. A well-trained model that captures informative predictors is robust and will continue to exhibit good results even when repeatedly challenged with new data. The external validation in the current study; thus, provides an assurance that our developed decision tree models passing this step are more likely domain interpretable[35].

Although the methodology by Syed et al. is innovative and follows a clear behavioral logic, we speculate that the signal shape of the raw acceleration right before and right after a candidate non-wear episode may be dependent on population age. Methodologies designed to identify non-wear time by the absence of acceleration are by definition independent of population characteristics as the collected data during non-wear is zero movement. The method proposed by Syed et al. characterizes non-wear by identifying the acceleration signal shape that marks the beginning and end of a non-wear episode. Thus, analyzing the physical activity behavior of the population and not the absence of acceleration, which might be more dependent on population characteristics. Thus, the poor performance seen in the present study by the CNN model by Syed et al. may be due to variations in the dataset populations. The CNN model was trained on data from an older population aged 40–84 years (mean = 62.74; SD = 10.25) compared to the datasets employed in the present study, aged 8.1–17.9 years (mean = 12.14, SD = 2.40) from the hip and thigh data, and aged 14.5–16.4 years old (mean = 15.4, SD = 0.37 years) from the wrist data. This is supported by our results showing that the *sunda_RF* model performs acceptable when identifying non-wear episodes shorter than 60 min on thigh and wrist data whereas *syed_CNN* performs poorly across all wear locations. This indicates that the model by Sundararajan et al. seem to be population characteristic agnostic as expected in contrast to the *syed_CNN* model. Finally, it is important to note that the *syed_CNN* model was originally trained on data with a frequency of 100 Hz. In our current study, we have used the model on data with frequencies of 50 Hz and 25 Hz. It is not clear if this difference in data frequency has had any effect on the model's performance. However, as movement frequencies is generally below 5 Hz we are confident that the 25 Hz data is sufficient to capture the true movement behavior for the subject mounting or unmounting the device.

It is customary practice to report performance of a model on a test split from the data that was used to train the model. This is called *internal validation* and is a sensible approach, but this also entails a minimum of variation between the train and test data. While the reported metrics, such as sensitivity, specificity and/or accuracy for the classification of non-wear time, are reported to be remarkably high by Syed et al.[18] and Sundararajan et al.[17] these results are obtained via cross-validation without an external validation dataset, which weakens the confidence of the generalizability of these models. Highly flexible models potentially overfit the training data distribution when independent test sets are not used or are prone to learn dataset-specific artifacts rather than more generalizable behavioral characteristics. As a rule of thumb, every machine learning and deep learning approach would benefit from larger training datasets, although we acknowledge that this is rarely practically

feasible. However, this may be particularly true for the model developed by Syed et al. Thus, to make their methodology more robust, the training data employed would benefit from a more diverse population and a higher number of participants as the varying signal shapes related to the mounting and un-mounting of the device may be dependent across age groups and other population characteristics. Therefore, for future developments within this area, we encourage the practice of validating performance on independent external datasets prior to publishing the model.

The main strength of the study is the use of external validation, as this is considered good evidence of method generalizability. One limitation is the construction of our ground truth datasets. As no accepted gold-standard exists within this field of research, ground truth estimates of non-wear vary across studies, thus hindering our ability to compare performance metrics. However, by utilizing raw accelerometer data, our methods are transparent, and no intermediate steps of the data collection and analysis are proprietary. Moreover, the results are based on a population of children and adolescents, and as such, cannot be generalized to older populations. Finally, other machine learning algorithms may be preferable, however; we chose to build a decision tree model as the complexity of this models still makes it possible to make meaningful interpretations. More research is needed to determine the effectiveness of other methods, i.e., logistic regression, gradient boosting, support vector machines and others.

Conclusions

In this study we present results on the performance and generalizability of existing methods and newly developed decision tree models for the classification of non-wear in free-living accelerometer recordings. Although the current available heuristic methods have shown promising results, they are subject to obvious limitation whereas recent complex machine learning methods may be prone to over-fitting as our results indicate. Furthermore, the quantity and quality of data are essential when training a machine learning model for a simple binary classification problem when researchers want to be able to generalize to other types of data. For this, we encourage the use of external validation to temper overoptimistic expectations of model performance in unseen data. Furthermore, for the crucial first step when analyzing accelerometer data (i.e., detecting non-wear time), we advise researchers to carefully select a proper method for this task.

Data availability

The classification models developed in this paper are available on request to the corresponding author as well as the raw data from the PHASAR study and the raw data from the in-house validation study.

Received: 9 June 2022; Accepted: 8 February 2023
Published online: 13 February 2023

References

1. Dowd, K. P. *et al.* A systematic literature review of reviews on techniques for physical activity measurement in adults: A DEDIPAC study. *Int. J. Behav. Nutr. Phys. Act.* **15**(1), 15. https://doi.org/10.1186/s12966-017-0636-2 (2018).
2. Loyen, A. *et al.* Sedentary time and physical activity surveillance through accelerometer pooling in four European countries. *Sports Med.* **47**(7), 1421–1435. https://doi.org/10.1007/s40279-016-0658-y (2017).
3. Montoye, A. H. K. *et al.* Raw and count data comparability of hip-worn actiGraph GT3X+ and link accelerometers. *Med. Sci. Sports Exerc.* **50**(5), 1103–1112. https://doi.org/10.1249/MSS.0000000000001534 (2018).
4. Migueles, J. H. *et al.* Comparability of accelerometer signal aggregation metrics across placements and dominant wrist cut points for the assessment of physical activity in adults. *Sci. Rep.* **9**(1), 18235. https://doi.org/10.1038/s41598-019-54267-y (2019).
5. Ae Lee, J. & Gill, J. Missing value imputation for physical activity data measured by accelerometer. *Stat. Methods Med. Res.* **27**(2), 490–506. https://doi.org/10.1177/0962280216633248 (2018).
6. Ainsworth, B. E. *et al.* Recommendations to improve the accuracy of estimates of physical activity derived from self report. *J. Phys. Act. Health.* **9**(Suppl 1 (0 1)), S76-84. https://doi.org/10.1123/jpah.9.s1.s76 (2012).
7. Hecht, A., Ma, S., Porszasz, J., Casaburi, R., for the COPD Clinical Research Network. Methodology for using long-term accelerometry monitoring to describe daily activity patterns in COPD. *COPD J. Chronic. Obstr. Pulm. Dis.* **6**(2), 121–129. https://doi.org/10.1080/15412550902755044 (2009).
8. Ruiz, J. R. *et al.* Objectively measured physical activity and sedentary time in european adolescents: The HELENA study. *Am. J. Epidemiol.* **174**(2), 173–184. https://doi.org/10.1093/aje/kwr068 (2011).
9. Troiano, R. P. *et al.* Physical activity in the united states measured by accelerometer. *Med. Sci. Sports Exerc.* **40**(1), 181–188. https://doi.org/10.1249/mss.0b013e31815a51b3 (2008).
10. Aadland, E., Andersen, L. B., Anderssen, S. A. & Resaland, G. K. A comparison of 10 accelerometer non-wear time criteria and logbooks in children. *BMC Public Health* **18**(1), 323. https://doi.org/10.1186/s12889-018-5212-4 (2018).
11. Toftager, M. *et al.* Accelerometer data reduction in adolescents: Effects on sample retention and bias. *Int. J. Behav. Nutr. Phys. Act.* **10**(1), 140. https://doi.org/10.1186/1479-5868-10-140 (2013).
12. van Hees, V. T. *et al.* Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. *PLoS ONE* **6**(7), e22922. https://doi.org/10.1371/journal.pone.0022922 (2011).
13. van Hees, V. T. *et al.* Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLoS ONE* **8**(4), e61691. https://doi.org/10.1371/journal.pone.0061691 (2013).
14. Duncan, S. *et al.* Wear-time compliance with a dual-accelerometer system for capturing 24-h behavioural profiles in children and adults. *Int. J. Environ. Res. Public Health.* **15**(7), E1296. https://doi.org/10.3390/ijerph15071296 (2018).
15. Rasmussen, M. G. B. *et al.* Short-term efficacy of reducing screen media use on physical activity, sleep, and physiological stress in families with children aged 4–14: Study protocol for the SCREENS randomized controlled trial. *BMC Public Health* **20**(1), 380. https://doi.org/10.1186/s12889-020-8458-6 (2020).
16. Zhou, S. M. *et al.* Classification of accelerometer wear and non-wear events in seconds for monitoring free-living physical activity. *BMJ Open* **5**(5), e007447. https://doi.org/10.1136/bmjopen-2014-007447 (2015).
17. Sundararajan, K. *et al.* Sleep classification from wrist-worn accelerometer data using random forests. *Sci. Rep.* **11**(1), 24. https://doi.org/10.1038/s41598-020-79217-x (2021).
18. Syed, S., Morseth, B., Hopstock, L. A. & Horsch, A. A novel algorithm to detect non-wear time from raw accelerometer data using deep convolutional neural networks. *Sci. Rep.* **11**(1), 8832. https://doi.org/10.1038/s41598-021-87757-z (2021).

19. Choi, L., Liu, Z., Matthews, C. E. & Buchowski, M. S. Validation of accelerometer wear and nonwear time classification algorithm. *Med. Sci. Sports Exerc.* **43**(2), 357–364. https://doi.org/10.1249/MSS.0b013e3181ed61a3 (2011).
20. Syed, S., Morseth, B., Hopstock, L. A. & Horsch, A. Evaluating the performance of raw and epoch non-wear algorithms using multiple accelerometers and electrocardiogram recordings. *Sci. Rep.* **10**(1), 5866. https://doi.org/10.1038/s41598-020-62821-2 (2020).
21. Pedersen, N. H. *et al.* Protocol for evaluating the impact of a national school policy on physical activity levels in Danish children and adolescents: the PHASAR study - a natural experiment. *BMC Public Health* **18**(1), 1245. https://doi.org/10.1186/s12889-018-6144-8 (2018).
22. Jaeschke, L., Steinbrecher, A., Jeran, S., Konigorski, S. & Pischon, T. Variability and reliability study of overall physical activity and activity intensity levels using 24 h-accelerometry-assessed data. *BMC Public Health* **18**(1), 530. https://doi.org/10.1186/s12889-018-5415-8 (2018).
23. Skovgaard, E. L., Pedersen, J., Møller, N. C., Grøntved, A. & Brønd, J. C. Manual annotation of time in bed using free-living recordings of accelerometry data. *Sensors.* **21**(24), 8442. https://doi.org/10.3390/s21248442 (2021).
24. Kuhn M, Wickham H. Tidymodels. Published online 2020. https://www.tidymodels.org.
25. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees. Published online 2019. https://CRAN.R-project.org/package=rpart.
26. Hutto, B. *et al.* Identifying accelerometer nonwear and wear time in older adults. *Int. J. Behav. Nutr. Phys. Act.* **10**(1), 1–8. https://doi.org/10.1186/1479-5868-10-120 (2013).
27. Doherty, A. *et al.* Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. *PLoS ONE* **12**(2), e0169649. https://doi.org/10.1371/journal.pone.0169649 (2017).
28. Knaier, R., Höchsmann, C., Infanger, D., Hinrichs, T. & Schmidt-Trucksäss, A. Validation of automatic wear-time detection algorithms in a free-living setting of wrist-worn and hip-worn ActiGraph GT3X. *BMC Public Health* **19**(1), 244. https://doi.org/10.1186/s12889-019-6568-9 (2019).
29. Ahmadi, M. N., Nathan, N., Sutherland, R., Wolfenden, L. & Trost, S. G. Non-wear or sleep? Evaluation of five non-wear detection algorithms for raw accelerometer data. *J Sports Sci.* **38**(4), 399–404. https://doi.org/10.1080/02640414.2019.1703301 (2020).
30. Barouni, A. *et al.* Ambulatory sleep scoring using accelerometers—distinguishing between nonwear and sleep/wake states. *PeerJ* **8**, e8284. https://doi.org/10.7717/peerj.8284 (2020).
31. Cooper, A. R. *et al.* Objectively measured physical activity and sedentary time in youth: The International children's accelerometry database (ICAD). *Int. J. Behav. Nutr. Phys. Act.* **12**(1), 113. https://doi.org/10.1186/s12966-015-0274-5 (2015).
32. Kwon, S., Burns, T. L., Levy, S. M. & Janz, K. F. Breaks in sedentary time during childhood and adolescence: Iowa bone development study. *Med. Sci. Sports Exerc.* **44**(6), 1075–1080. https://doi.org/10.1249/MSS.0b013e318245ca20 (2012).
33. Vert, A. *et al.* Detecting accelerometer non-wear periods using change in acceleration combined with rate-of-change in temperature. *BMC Med. Res. Methodol.* **22**(1), 147. https://doi.org/10.1186/s12874-022-01633-6 (2022).
34. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005 (2016).
35. Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. M. Prognosis and prognostic research: Validating a prognostic model. *BMJ* **338**, b605. https://doi.org/10.1136/bmj.b605 (2009).

Acknowledgements

This work was supported by funding from TrygFonden (grant number ID 130081 and 115606) and the European Research Council (grant number 716657).

Author contributions

Study concept and design: E.L.S., A.G. and J.C.B.; Data collection: N.H.P., K.T.L, and M.A.R.; Analysis and interpretation of data: E.L.S. and J.C.B.; Drafting of the manuscript: E.L.S. and J.C.B.; Critical revision of the manuscript: All authors.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.L.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com

Appendix III

Improving Sleep Quality Estimation in Children and Adolescents: A Comparative Study of Machine Learning and Deep Learning Techniques Utilizing Free-Living Accelerometer Data from Thigh-Worn Devices and EEG-Based Sleep Tracking

This manuscript is under preparation for submission to **SLEEP**, the official journal of the Sleep Research Society (SRS).

Improving Sleep Quality Estimation in Children and Adolescents: A Comparative Study of Machine Learning and Deep Learning Techniques Utilizing Free-Living Accelerometer Data from Thigh-Worn Devices and EEG-Based Sleep Tracking

Esben Høegholm Lykke
University of Southern Denmark
eskovgaard@health.sdu.dk

Anders Grøntved
University of Southern Denmark
agroentved@health.sdu.dk

Peter Lund Kristensen
University of Southern Denmark
plkristensen@health.sdu.dk

Jesper Schmidt-Persson
University of Southern Denmark
jespedersen@health.sdu.dk

Sara Overgaard Sørensen
University of Southern Denmark
sosorensen@health.sdu.dk

Sofie Rath Mortensen
University of Southern Denmark
srmortensen@health.sdu.dk

Jan Christian Brønd
University of Southern Denmark
jbrond@health.sdu.dk

2023-08-30

Abstract

Accurate assessment of sleep is vital in sleep research, but the gold standard, polysomnography, is costly and impractical for large-scale studiesand multiple consecutive days of assessment. An affordable alternative is using wearable accelerometers. While wrist and hip-worn devices are commonly used in sleep research, thigh-worn accelerometers have been relatively unexplored. Our study evaluated machine learning and deep learning models utilizing data from thigh-worn accelerometers to estimate sleep and sleep quality metrics, comparing them with an EEG-based sleep monitor. The dataset consisted of data from 585 days and nights, comprising accelerometry and EEG-based sleep estimates from children aged 4-17 years. We employed both sequential and multiclass model strategies on both raw and filtered data. The most effective model was XGBoost, which performed well when applied to 5-minute median filtered data, exhibiting small mean differences (bias) in sleep period time (0.2 minutes), total sleep time (-7.0 minutes), sleep efficiency (-1.1%), and wake after sleep onset (-0.9 minutes). Furthermore, the XGBoost model showed a robust correlation (0.66, 95% CI: 0.61 - 0.7) with total sleep time, indicating its potential. However, despite these promising results in bias, our study revealed limits of agreements (e.g., total sleep time LoA (95%CI): -95.5 (-105.2;-88) minutes to 81.4 (72.4;92.5) minutes) in accordance with previous research on hip- and wrist-worn devices. In conclusion, we present promising results in using machine learning techniques to estimate sleep quality metrics on a group level, however, accurately classifying awake periods during in-bed time remained challenging. Moreover, additional improvements are necessary to accurately assess individual sleep quality metrics based on thigh-worn accelerometry data due to the notable limits of agreement.

I. Introduction

A vast body of research highlights the critical role of sleep in maintaining both mental and physical health[1–4]. Consequently, accurate sleep assessment methods are crucial for tracking sleep patterns and improving our understanding of the sleep-health relationship. Furthermore, the ease of use and high acceptability of methods to assess sleep are essential to facilitate large-scale, longitudinal studies.

The traditional gold standard for objective sleep measurement, laboratory-based polysomnography (PSG), has been found to be impractical in large-scale observational- and experimental studies due to its high cost, need for professional administration, and susceptibility to rater bias[5,6], although recent advances has been made to automate the scoring of PSG data[7]. As an alternative, diaries have been used due to their cost-effectiveness and simplicity, although they are subject to recall bias and other limitations[8]. An innovative approach involves device-based measurement methods. These tools, which estimates a number of sleep metrics including sleep duration, are advantageous due to their reduced participant burden and elimination of potential recall biases. A prominent example of such tools is body-worn accelerometers, which offer a practical and affordable means of objectively assessing sleep patterns at home for extended periods. Accelerometers collect continuous, high-resolution data for several weeks without requiring recharging, further minimizing participant burden. Their use in sleep and wake classification began with a wrist movement-based algorithm developed in 1982, and validated using PSG[9]. This algorithm was refined in 1992[10], leading to the widely adopted Cole-Kripke model. With advancements in the field, a variety of techniques, including heuristic algorithms, machine learning models, regression, and deep learning, are now used to analyze data from hip and wrist-worn accelerometers[10–15].

While wrist and hip-worn devices have benefited from extensive methodological development, thigh-worn accelerometers have not seen the same level of advancement. Existing studies mainly focus on distinguishing sleep from wakefulness, with emphasis on defining 'waking time' and 'bedtime'[16–19]. Recent strides in estimating sleep duration using thigh-worn devices have been made, including the introduction of a promising algorithm and its comparison against PSG[20]. Despite these advancements, the application of machine learning techniques in this area is still unexplored. Considering the potential of thigh-worn accelerometers for accurate physical behavior assessment[21–23], there is a significant research gap. Therefore, future studies need to develop techniques similar to those used for wrist and hip-worn accelerometers, with the ultimate goal of establishing a more holistic, accurate, and user-friendly method of sleep and physical activity tracking.

The Zmachine® Insight+ (ZM) emerges as a valuable tool within this landscape. Favorably validated against PSG[24,25], the ZM provides comparable data without the high costs or the need for professional monitoring typically associated with PSG. Crucially, the ZM facilitates multi-night analysis in free-living conditions due to its ease of use[26], capturing the natural variations in sleep patterns. This makes it advantageous over single-night PSG, particularly as a gold standard data source in machine learning tasks, as it provides multiple nights of measurements without inter-rater bias. Despite these benefits, the ZM, like PSG, still poses a significant participant burden and cost, reinforcing the need for more accessible alternatives like accelerometers.

Our primary objective in this study was to evaluate a range of machine learning and deep learning models, utilizing the raw data collected from a tri-axial thigh-worn accelerometer to estimate in-bed and sleep time. To ensure the reliability and effectiveness of our models, we compared their outputs with an electroencephalography-based (EEG) sleep tracking device, which we, in this current study, considered as the criterion measure for assessing sleep. Furthermore, our secondary goal was to assess the developed models' performance in evaluating important sleep quality metrics, including sleep period time (SPT), total sleep time (TST), sleep efficiency (SE), latency until persistent sleep (LPS), and wake after sleep onset (WASO).

II. Methods

i. Dataset and Participants

The current study leverages data from the SCREENS trial[27,28], a study conducted from June 2019 to March 2021 in the Region of Southern Denmark, Southern Denmark, that evaluated the effect of limiting screen media usage within Danish families. For our analysis, we focused on data from child participants aged between 4 and 17 years within the SCREENS cohort (mean age 9.1 years). Our primary sources of

data were accelerometer readings from Axivity AX3 devices attached to the children's thighs, and EEG-derived sleep states and sleep quality metrics from the ZM device. The Axivity AX3, an unobtrusive 3-axis accelerometer, was positioned midway between the hip and knee on the right anterior thigh, recording participant movement data.

Sleep state information was extracted using the ZM, a product of General Sleep Corporation. The ZM, which utilizes advanced EEG hardware and signal processing algorithms, employs three self-adhesive, disposable sensors placed outside the hairline for reliable EEG signal acquisition. The participants of the SCREENS study were instructed to attach the device when they went to bed and remove the device upon leaving the bed. The ZM uses two proprietary algorithms: Z-ALG and Z-PLUS. The Z-ALG is utilized for accurate sleep detection, showcasing its suitability for in-home monitoring[24], while the Z-PLUS effectively differentiates sleep stages, as evidenced by its alignment with expert evaluations using PSG data[25]. In the current study, we treated all sleep stages (light sleep (N1 & N2), deep sleep (N3), and REM sleep) as a single category effectively deducing the output of the ZM to "awake" and "asleep" as the ability to distinguish sleep stages are not a necessity to derive the sleep quality metrics of interest and to simplify the learning process of the machine learning algorithms.

Figure 1 illustrates the selection criteria applied to the children's recordings from the SCREENS study. Only ZM recordings accompanied by complete accelerometer data and lasting between 7 and 14 hours were considered. Nights during which the ZM reported sensor issues were excluded. Consequently, a total of 585 nights from 151 children were included in the study, with a mean of 3.87 nights per child (SD = 1.86). The children whose recordings were considered had an average age of 9.4 years, with a standard deviation of 2.1. The ZM predictions encompassed 696,779 epochs, each 30 seconds long. Notably, approximately 84% of the total ZM recording duration was classified as sleep, resulting in an imbalance of the classes during the nightly recordings.

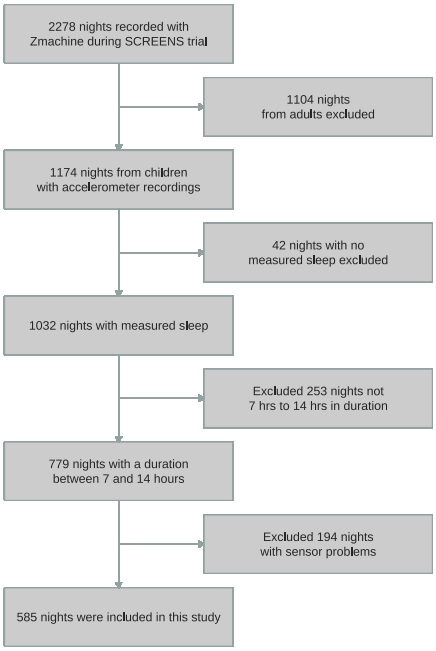

Figure 1: *Flowchart of eligible ZM recording nights included in the study.*

Finally, we affirm that the SCREENS study received approval from the Regional Scientific Commit-

tee of Southern Denmark, and all data handling processes complied with the General Data Protection Regulation (GDPR), ensuring the ethical and secure management of participant information.

ii. Data Preprocessing and Feature Extraction

In this study, data processing of the raw accelerometer data began with a low-pass filtration step using a 4th order Butterworth filter with a 5 Hz cut-off frequency to eliminate high-frequency noise similar to methods described by Skotte et al.[21]. Any non-wear data was removed using previously described methods[29] and data was resampled to 30-second epochs so every sample classified by the models corresponds to a 30-second epoch scored during the ZM recordings. Subsequently, we performed a feature extraction process that yielded a set of 64 features, providing a robust characterization of the data. Extracted from accelerometer and temperature signals, these features include temporal elements that use both lag and lead values, capturing dynamic data trends by incorporating measurements from preceding and upcoming epochs. Furthermore, inspired by Walch et al.[30], we incorporated sensor-independent features to encapsulate circadian rhythms. These features offer unique insights not directly discernible from sensor outputs and are meant to approximate the changing drive of the circadian clock to sleep over the course of the night (see Figure 2). Furthermore, the feature set was enriched by including signal characteristics, which encompass vector magnitude, mean crossing rate, skewness, and kurtosis for each of the x, y, and z dimensions. Subsequently, we merged the ZM and corresponding accelerometer recordings. Any overlapping time between the ZM and accelerometer data was treated as 'in-bed' time, with the remaining time considered 'out-of-bed'. This process yielded a dataset providing a around the clock temporal view of each participant's activity and sleep patterns.

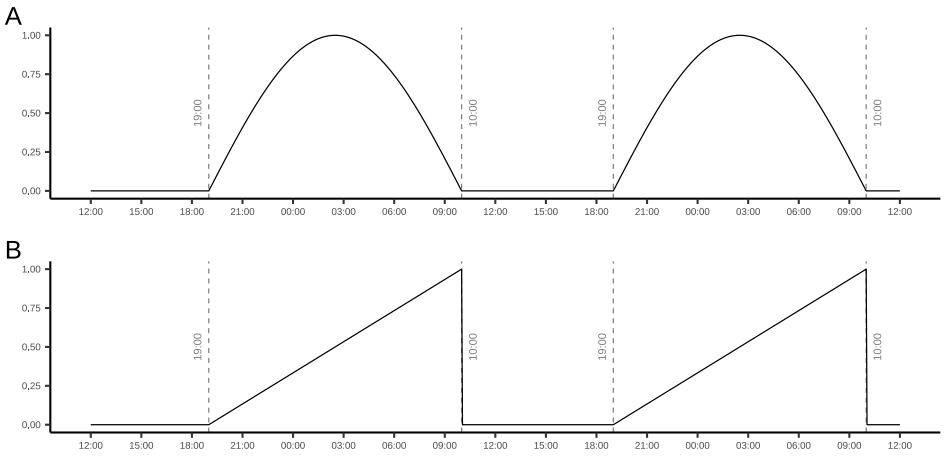

Figure 2: *Sensor-independent features of circadian rhythms across two consecutive nights. A) cosinus feature, B) linear feature.*

In addition to the engineered features, we chose to incorporate the median-filtered raw predictions from the ZM device into our modeling process. This choice was informed by the understanding that children typically experience around five to eight sleep cycles per night, with awakenings most likely at the end of each cycle[31]. In examining the raw ZM predictions, we observed a notable overestimation in the number of awakenings for the children in our study, surpassing expected counts based on typical sleep cycle patterns (refer to Figure 3). The average sleep efficiency determined by the ZM for our sample was 83%, which falls slightly below the recognized standards—85% is seen as good efficiency, and above 90% as ideal. In fact, prior research has indicated sleep efficiencies of over 90% in similar child cohorts[32,33]. This discrepancy suggests that the raw ZM predictions might be overestimating awake periods. many of these brief awakenings could be considered as noise, which when present in the data, can potentially hinder the learning process of machine learning algorithms by obscuring the underlying patterns that the

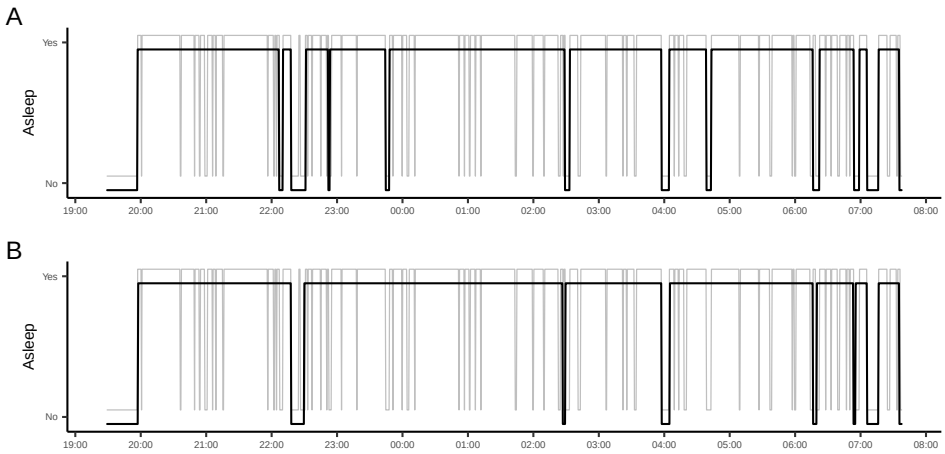

Figure 3: *The difference in number of awakenings between the raw ZM predictions vs. 5-minute, and 10-minute median filtered predictions for a random night (boy, 9 years). Grey line is the raw predictions, black line is the median filtered predictions. A: 5-minute median filter on raw ZM predictions, B: 10-minute median filter on raw ZM predictions.*

algorithms are trying to learn, leading to less accurate predictions. Consequently, we elected to train and evaluate our models using not only the raw ZM output, but also versions that were subjected to 5-minute and 10-minute median filters. This approach, by mitigating this noise, resulted in an anticipated, more age-appropriate count of awakenings per night, which to the best of our knowledge provided a more accurate depiction of children's sleep patterns (see Table 1).

iii. Algorithms

We employed two different model strategies to assess sleep patterns from thigh-mounted accelerometer data. The first model strategy was designed as a sequence of two models, each functioning as a binary classifier. This approach aimed to simplify the prediction task by decomposing the multiclass problem of classifying 'out-of-bed-awake', 'in-bed-awake', and 'in-bed-asleep' into two binary stages: first predicting 'in-bed' time, then 'sleep' time. The output from the first set of binary classifiers, which predicted in-bed time, was subjected to a 5-minute median filter to remove transient in-bed time blips. This process enabled us to establish a single continuous time interval that we identified as the SPT, the total time spent in bed attempting to sleep. The SPT then served as the input for the second stage of binary classifiers in the sequence, further enhancing their predictive accuracy for sleep time. We applied this sequential strategy using the following four machine learning algorithms:

1. Logistic Regression: Logistic regression served as a simple and fast baseline model. However, due to its linear nature, it may struggle with capturing complex relationships and non-linear patterns present in the accelerometer data.

2. Decision Tree: Decision trees are capable of handling non-linear patterns and are easily interpretable. However, they are prone to overfitting, particularly when dealing with complex patterns that require simultaneous consideration of multiple features. To combat this, we used a maximum tree depth of 8.

3. Single-layer Feed-forward Neural Network: Single-layer feed-forward neural networks can effectively capture non-linear relationships, even with their relatively simple structure. However, they tend to be more challenging to interpret compared to simpler models. Additionally, careful tuning of the network's architecture and training process is required to mitigate the risk of overfitting.

4. XGBoost: XGBoost is a powerful algorithm known for its ability to provide highly accurate predictions and handle complex, non-linear patterns in the data. It also incorporates built-in methods to prevent overfitting. However, training XGBoost models can be computationally intensive, and interpreting the predictions it generates can pose challenges.

In parallel, we also employed a multiclass algorithm as the second model strategy using a bidirectional Long Short-Term Memory (biLSTM)[34] neural network which also incorporates temporal aspects of the data. This network, which was designed to predict three distinct classes: 'out-of-bed-awake', 'in-bed-awake', and 'in-bed-asleep', was configured with four layers and 128 hidden units per layer. This balance between model complexity and training efficiency was intended to facilitate learning of intricate patterns while ensuring feasible training times. The bidirectional nature of the LSTM enhanced data interpretation and reduced overfitting by doubling the hidden units at each time step. The LSTM model used sequences of tensors as input, with each sequence spanning 10 minutes and a step size of one epoch. As demonstrated by previous studies such as those by Sano et al.[35] and Chen et al.[36], LSTM models have shown great promise in sleep detection using accelerometer data, thanks to their ability to capture complex temporal patterns.

iv. Model Training

We trained four pairs of models in sequence, with each pair distinguishing between in-bed/out-of-bed and asleep/awake states, respectively. We divided our dataset randomly into a training set and a testing set, with each containing roughly half of the subjects. The splitting of the data was ensured to not have samples from the same night simultaneously present in both sets. To optimize hyperparameters, we performed a 10-fold Monte Carlo cross-validation on a regular grid (i.e., for each hyperparameter, a range of values at evenly-spaced intervals was selected) comprising 20 different combinations of hyperparameters. The F1 score served as the optimization metric. The best-performing set of hyperparameters was then used to fit the models to the full training dataset. This approach allowed us to maximize performance by leveraging all available training data to estimate the model parameters. An imbalance was observed with the in-bed time determined in the initial step of the sequential model strategy which after extracting the in-bed time from the initial sequential models, the imbalance on the resulting dataset could cause biases during model training, as models favor predicting the majority class. To account for this imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE)[37]. SMOTE generates new samples by interpolating random samples with their nearest neighbors. We utilized the themis R package[38] to implement SMOTE, resulting in a balanced distribution of training samples across both classes.

The biLSTM model was trained to differentiate between three classes: out-of-bed-awake, in-bed-awake, and in-bed-asleep. The data used for training the biLSTM was randomly divided into training, validation, and test sets, based on a 50/25/25 split. Again, we ensured that data from the same night was not present across different sets. The model was trained using the Adam optimizer, selected for its computational efficiency and adaptability of the learning rate during training. Given the multiclass classification task with mutually exclusive classes, we employed the cross-entropy loss function. To obtain a probability distribution over the classes, the softmax activation function was applied at the output layer. We evaluated the model's performance using the F1 score on both the training and validation sets. We implemented early stopping with a patience of 3 epochs, halting the training process if there was no improvement in the validation loss over three consecutive epochs.

v. Model Validation

In our study, we utilized standard evaluation metrics to assess the performance of each model on an epoch-to-epoch basis. These include

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$precision = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$F_1 = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity}$$

where NPV is negative predictive value, F_1 is the F1 score, TP is true positives, FP is false positives, TN is true negatives, and FN is false negatives.

In the context of our sequential model strategy, the initial models were tasked with the binary classification of in-bed vs. out-of-bed. For this task, we assessed performance using the F1-score, accuracy, sensitivity, specificity, and precision metrics. The second models in our sequential model strategy focused on the binary classification of asleep vs. awake. For these models, we considered the same metrics, in addition to the negative predictive rate. The class imbalance in this case led us to compute the F1 score as an unweighted macro-average. Additionally, we evaluated the multiclass classifier, biLSTM, using the same metrics. To do this, we considered the multiclass output as to binary classifications, where the first was out-of-bed vs the rest and the second binary classification as in-bed-awake vs in-bed-asleep. To further illustrate model performance, we provide confusion matrices for the full dataset, encompassing both in-bed and out-of-bed data. These matrices report relative counts, column percentages (the proportion of the true class accurately predicted), and row percentages (the proportion of predictions correctly classified). We considered both the in-bed/out-of-bed and awake/asleep scoring tasks as binary classification problems, designating in-bed and asleep as the positive labels and out-of-bed and awake as the negative labels in accordance with previous research[39,40].

To assess the performance of our models in deriving sleep quality metrics, we utilized Bland-Altman plots and Pearson correlations. The Bland-Altman method was employed specifically to determine the level of agreement between two measurement techniques. Given the nature of our dataset, which contains multiple observations per subject but not necessarily equal number of observations, we employed a bootstrap procedure to account for this added variability. We first calculated the mean difference (bias) and then defined the limits of agreement (LOA) as the mean difference plus or minus 1.96 times the standard deviation of these differences. Acknowledging the possibility of non-normality and potential skewness in our data, we chose to apply a bias-corrected and accelerated (BCa) bootstrap method[41]. This approach allowed us to better address potential bias in our estimates and the inherent intra-subject variability. Utilizing 10,000 bootstrap replicates, we estimated the 95% confidence intervals for both the bias and the LOA, thus ensuring robustness in our measurements. The sleep quality metrics included are defined as follows in accordance with the ZM definitions:

1. Sleep Period Time (SPT) - This refers to the total duration of time in bed with the intention to sleep, which is defined as the time from the start to the end of the ZM recording.

2. Total Sleep Time (TST) - This is the time spent asleep within the SPT.

3. Sleep Efficiency (SE) - This is the ratio between TST and SPT, representing the proportion of the sleep period that was actually spent asleep.

4. Latency Until Persistent Sleep (LPS) - This metric represents the time it takes to transition from wakefulness to sustained sleep. It is calculated as the time from the beginning of the ZM recording until the first period when 10 out of 12 minutes are scored as sleep.

5. Wake After Sleep Onset (WASO) - This refers to the time spent awake after initially falling asleep and before the final awakening. In our analysis, a period is counted as 'awake' only if it consists of 3 or more contiguous 30-second epochs which is also how the ZM summarizes WASO.

R version 4.3.0 (2023-04-21)[42] and the Tidymodels[43] and Tidyverse[44] suite of packages were used as the core tools for model development and analyses. Python version 3.10.6[45] and PyTorch[46] were used to implement the biLSTM model.

III. Results

As reported in Table 1 the sleep quality metrics derived from ZM predictions were modified by the implementation of 5-minute and 10-minute median filters. SPT were consistent across raw and filtered datasets (mean: 9.2 ± 2.1 hours), corresponding to the length of the ZM recording. TST and SE increased in the filtered data, implying the filters categorize some wakefulness as sleep. Specifically, TST increased from a raw mean of 7.7 ± 1.9 hours to 8.1 ± 2.0 hours (5-minute filter) and 8.2 ± 2.1 hours (10-minute filter), while SE rose from 82.6 ± 12.0% to 86.4 ± 12.7% and 87.5 ± 12.9% respectively. LPS also increased, suggesting the filter removes brief awakenings at sleep onset, leading to a prolonged time to persistent sleep. A change was seen in WASO, which dropped from 39.0 ± 33.6 minutes in raw data to 30.6 ± 46.8 minutes and 22.3 ± 55.4 minutes in the 5-minute and 10-minute filtered data, respectively. The number of awakenings was also considerably reduced with the application of filters. In the raw data, the average number of awakenings was 34.46 ± 11.33 per night, which reduced to 4.43 ± 3.26 and 1.95 ± 2.01 for the 5-minute and 10-minute filtered data sets respectively.

Table 1: *Overview of characteristics of the ZM sleep quality summaries per night (585 nights from 151 children). Values are represented as mean (SD). Hrs: hours, min: minutes.*

	SPT (hrs)	TST (hrs)	SE (%)	LPS (min)	WASO (min)	Awakenings (N)
Raw ZM Predictions	9.2 (2.1)	7.7 (1.9)	82.6 (12)	34.5 (27.9)	39 (33.6)	34.5 (11.3)
5-Min Median	9.2 (2.1)	8.1 (2)	86.4 (12.7)	36.3 (39.8)	30.6 (46.8)	4.4 (3.3)
10-Min Median	9.2 (2.1)	8.2 (2.1)	87.5 (12.9)	38 (48.7)	22.3 (55.4)	1.9 (2)

i. Performance on Epoch-to-Epoch Basis

The epoch-to-epoch evaluation of predicting in-bed time is outlined in Table 2, and demonstrates practically equivalent performance across all model types. The F1 score ranges from 94.4% (Decision Tree) to 95.4% (XGBoost), while accuracy ranges from 95.3% (Decision Tree) to 96.1% (XGBoost). Sensitivity, Precision, and Specificity also demonstrate consistent results across the different models. The XGBoost model provide the best performance with an F1 score of 95.4% and accuracy of 96.1%, although only outpacing the other models marginally.

Table 2: *Performance metrics of the classification of in-bed/out-of-bed time of the included models.*

	F1 Score (%)	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)
Decision Tree	94.4	95.3	93.1	95.6	96.9
Logistic Regression	95.0	95.7	95.0	94.9	96.3
Feed-Forward Neural Net	95.0	95.8	95.1	95.0	96.3
XGBoost	95.4	96.1	95.8	94.9	96.2
biLSTM	95.2	95.3	95.3	95.1	95.3

Table 3 details the performance of all sequential model types on raw and median-filtered (5 and 10 minute) ZM predictions for sleep/wake classification. For raw ZM predictions, the F1 scores, which are unweighted macro averages, range from 65.6% (biLSTM) to 76.2% (XGBoost). All models perform comparably, but the low specificity values (62.5% to 70.9%) suggest difficulty in correctly classifying awake epochs. Applying a 5-minute median filter improves the performance metrics. The XGBoost model tops the charts with an F1 score of 79.2% and NPV of 74.0%. However, specificity still remains low, with values between 54.7% (XGBoost) and 74.8% (Logistic Regression) across all models. With a 10-minute median filter, the metrics improve further. The XGBoost model still leads with an F1 score of 80.9% and an NPV of 75.9%. But, specificity remains low, ranging from 57.5% (Decision Tree) to 76.4% (Logistic Regression) across all models.

Table 3: *Performance metrics of the sleep/wake classification of the included models.*

	F1 Score (%)	Precision (%)	NPV (%)	Sensitivity (%)	Specificity (%)
Raw ZM Predictions					
Decision Tree	72.9	93.2	48.4	86.3	67.1
Logistic Regression	71.0	93.7	43.9	82.7	70.9

Neural Network	71.8	93.8	45.1	83.6	70.8
XGBoost	76.2	92.8	58.0	91.3	62.8
biLSTM	65.6	80.6	80.6	62.5	62.5
5-Min Median					
Decision Tree	75.5	94.2	55.5	93.4	59.0
Logistic Regression	68.3	95.8	36.0	81.4	74.8
Neural Network	71.7	95.8	41.6	85.6	73.1
XGBoost	79.2	93.9	74.0	97.3	54.7
biLSTM	70.3	84.6	84.6	66.2	66.2
10-Min Median					
Decision Tree	76.3	94.7	58.1	94.9	57.5
Logistic Regression	68.0	96.5	34.3	81.9	76.4
Neural Network	71.0	96.1	39.5	86.5	71.4
XGBoost	80.9	94.9	75.8	97.7	57.6
biLSTM	70.9	75.1	75.1	68.5	68.5

A complete set of confusion matrices generated from data both containing the out-of-bed and in-bed time are presented in Figure 4. These matrices showcase the epoch-to-epoch performance of all sequential models in distinguishing between 'awake' and 'asleep' states, regardless of whether the subject is 'in-bed' or 'out-of-bed'. However, it's important to note that the binary nature of these sequential models means they cannot provide direct information about the classification of the 'in-bed-awake' state. In contrast, the biLSTM model, which also categorizes the 'in-bed-awake' state as a distinct class, appears to have less success in classifying this particular state.

ii. Evaluation of sleep quality metrics

Table 4 presents a comparative analysis of the included models used to predict various sleep quality metrics (SPT, TST, SE, LPS, WASO) using the 5-minute median filtered ZM predictions. To see the full table including models developed from raw ZM predictions and 10-minute median filtered ZM predictions, see table 1 in supplementary materials. In terms of bias, the decision tree model consistently underestimated SPT, TST, and SE, and overestimated LPS and WASO in comparison to ZM. The logistic regression model had similar trends, with more pronounced underestimation in TST and overestimation in LPS. The feed-forward neural network also exhibited similar bias as the decision tree and the logistic regression models, but with a higher overestimation in WASO. On the other hand, the XGBoost model showed least bias among all, especially in its 5-minute median predictions. Considering LOA, the decision tree had higher variability in the differences across different sleep quality metrics and filtering techniques, particularly for LPS and WASO, which indicates lower agreement with ZM. Other models had comparable LOA but with notable exceptions. For example, TST LOA for the logistic regression model was particularly wide in the 5-minute median predictions. Correlation-wise, the pearson coefficient, revealed that the XGBoost model consistently had the highest correlation with ZM across all sleep quality metrics and filtering methods Notably, the XGBoost's 5-minute median predictions showed the strongest correlation (0.66) for TST among all models and filtering techniques.

Table 4: *Summary of bias, limits of agreement, and Pearson correlation for various sleep parameter predictions (SPT, TST,SE, LPS, WASO) using different machine learning and deep learning models (decision tree, logistic regression, feed-forward neural network, XGBoost) on raw ZM predictions, 5-minute and 10-minute median predictions. Each value is provided with its 95% confidence interval (CI).*

	Bias (95% CI)	Lower LOA (95% CI)	Upper LOA (95% CI)	Pearson, r (95% CI)
5-Min Median - Decision Tree				
SPT (min)	-21.6 (-25.6;-17.6)	-117.5 (-125.6;-110.7)	74.2 (63.9;85.9)	0.54 (0.48;0.6)
TST (min)	-50.5 (-55.2;-46)	-161.4 (-175.8;-151.3)	60.4 (51.5;71.7)	0.48 (0.42;0.54)
SE (%)	-5.5 (-6.3;-4.7)	-23.9 (-26.4;-22.2)	12.9 (11.6;14.6)	0.22 (0.14;0.29)
LPS (min)	24.6 (19.7;29.1)	-88.8 (-115;-77.3)	138 (126.2;156.7)	0.06 (-0.02;0.14)
WASO (min)	9.9 (6.5;14)	-79.4 (-109;-63.1)	99.2 (80;136.1)	0.15 (0.07;0.22)
5-Min Median - Logistic Regression				
SPT (min)	-3.7 (-8;1)	-112.2 (-120.9;-105.2)	104.8 (94;117.4)	0.38 (0.3;0.44)
TST (min)	-139.7 (-146.9;-133)	-305.6 (-323.6;-291.8)	26.2 (16.1;38.6)	0.09 (0.01;0.17)

SE (%)	-23.2 (-24.3;-22.2)	-48.1 (-50.9;-46.1)	1.7 (0.1;3.8)	0.13 (0.05;0.21)
LPS (min)	58.1 (53.4;62.6)	-52.3 (-75;-40.1)	168.6 (155.9;187.7)	0.05 (-0.03;0.13)
WASO (min)	45.4 (41.7;49.7)	-50.7 (-74.4;-38.4)	141.5 (126.8;173)	0.19 (0.11;0.27)
5-Min Median - Feed-Forward Neural Net				
SPT (min)	-3.9 (-8.1;0.9)	-112.7 (-122;-105.2)	104.9 (94.1;118.4)	0.38 (0.3;0.44)
TST (min)	-126.5 (-132.8;-120.3)	-276.8 (-291.3;-264.7)	23.9 (14.8;33.9)	0.25 (0.17;0.32)
SE (%)	-20.9 (-21.9;-19.9)	-44.3 (-46.3;-42.5)	2.5 (1.1;4)	0.21 (0.13;0.29)
LPS (min)	35.3 (30.7;39.8)	-75.8 (-102.3;-63.4)	146.5 (134.4;166.9)	0.07 (-0.01;0.15)
WASO (min)	45 (41.2;49.2)	-51.8 (-76.4;-39.1)	141.7 (125.8;174.1)	0.21 (0.14;0.29)
5-Min Median - XGboost				
SPT (min)	0.2 (-3.7;4.5)	-97.4 (-106.2;-90.3)	97.8 (86.6;111)	0.56 (0.5;0.61)
TST (min)	-7 (-10.8;-3.3)	-95.5 (-105.2;-88)	81.4 (72.4;92.5)	0.66 (0.61;0.7)
SE (%)	-1.1 (-1.7;-0.5)	-15.6 (-17;-14.4)	13.3 (12.2;14.7)	0.44 (0.38;0.51)
LPS (min)	28.5 (23.9;32.6)	-76.4 (-104.2;-63.3)	133.4 (120.4;154.2)	0.12 (0.04;0.2)
WASO (min)	-0.9 (-3.9;3)	-83.4 (-113.1;-66)	81.7 (62;119.6)	0.26 (0.18;0.33)
5-Min Median - biLSTM				
SPT (min)	-36.1 (-41.7;-30)	-136.1 (-146.3;-126.9)	64 (51.1;78.6)	0.54 (0.45;0.62)
TST (min)	12.8 (7.4;18.3)	-80.1 (-89.8;-72.3)	105.8 (94.3;118.8)	0.63 (0.55;0.69)
SE (%)	8 (7.2;8.8)	-5.1 (-6.8;-3.8)	21.1 (19.5;23.1)	0.16 (0.04;0.27)
LPS (min)	-15.7 (-25.9;-7.5)	-169 (-230.7;-127.9)	137.6 (101.1;184.9)	0.09 (-0.02;0.2)
WASO (min)	-3 (-9.9;7.7)	-144.1 (-197.2;-107.2)	138.1 (90.8;211.4)	0.02 (-0.1;0.13)

Figure 5 shows the agreement between the XGBoost model, trained on 5-minute median filtered ZM predictions, and the 5-minute median-smoothed ZM-derived sleep quality metrics. The Bland-Altman plot for the SPT and TST indicates a minimal average difference with the ZM, as evidenced by a bias close to zero. The scatterplot for SPT also demonstrates a positive trend, indicating a moderate linear correlation between the XGBoost model and the ZM-derived sleep quality metrics. The bias and LOA for TST are comparable to those observed for SPT, indicating a consistent level of agreement between the two methods. The scatterplot for TST also shows a slightly higher correlation, primarily driven by the absence of extreme outliers.Furthermore, the remaining three sleep quality metrics, SE, LPS, and WASO, exhibit heteroscedasticity in contrast to SPT and TST. A moderate positive linear correlation is observed between the XGBoost model and ZM-derived sleep quality metrics for SE, however, a poor correlation is observed for LPS and WASO.

IV. Discussion

To select the most optimal method for estimating sleep from thigh-worn accelerometers, we evaluated various models for predicting in-bed and sleep time and their derived sleep quality metrics. We trained and evaluated the models using raw and median-filtered sleep estimates from the ZM EEG-based sleep monitor. In general, all sequential models performed well at predicting in-bed time. More challenging was it to distinguish wake from sleep on the extracted in-bed time. Moreover, even though the multiclass biLSTM showed good performance across F1 score, precision and NPV, the derived sleep quality metrics were not on par with the XGBoost model which demonstrated the highest performance metrics across all evaluations, including epoch-to-epoch prediction and all sleep quality metrics. Despite this, all models showed low specificity values, indicating difficulty in correctly classifying awake epochs during time in bed. The application of 5-minute and 10-minute median filters improved the performance metrics of all models. Median filtering increase total sleep time and sleep efficiency, while reducing wake after sleep onset and the number of awakenings. The XGBoost model provided the smallest bias and highest correlation with all ZM sleep quality metrics.

Limited research exists regarding the epoch-to-epoch effectiveness of classifying in-bed time based on data from thigh-worn accelerometers. Nevertheless, Carlson and colleagues provided compelling insights. They demonstrated that a third-party algorithm, "ProcessingPal," and a proprietary one, "CREA," achieved accuracies of 91% and 86% respectively. These algorithms, evaluated against self-reported measures among adolescents and adults[16], produced F1 scores as high as 95% and 96%. These figures are consistent with the performance of our sequential models, which also achieved F1 scores and accuracy scores exceeding 95% in identifying in-bed time. In our study, in-bed time is equated with SPT. All models, with the exception of XGBoost, underestimated SPT. The biLSTM model showed the

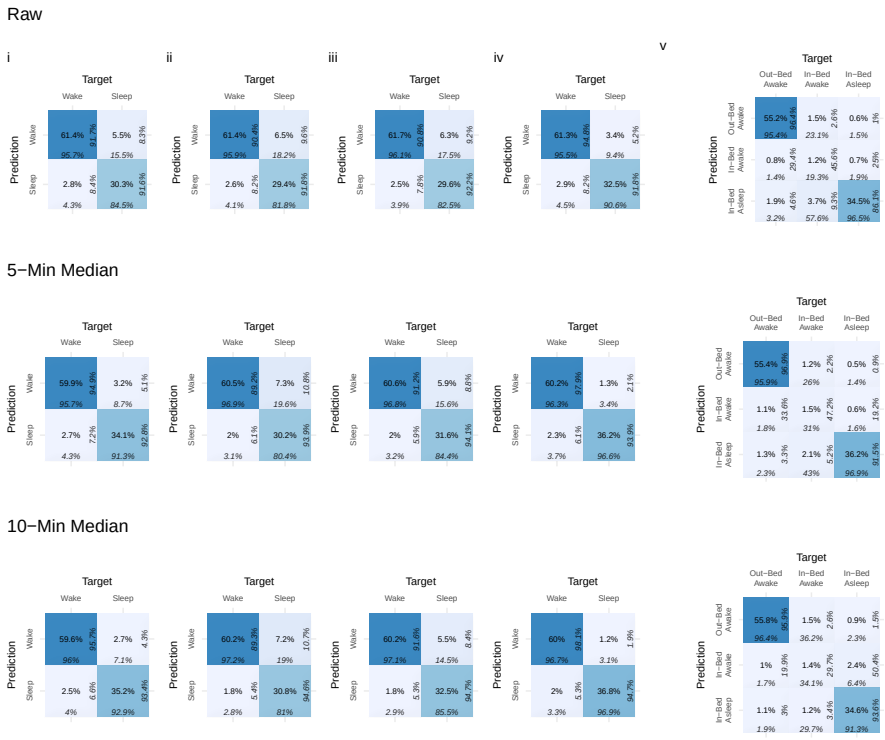

Figure 4: *Confusion matrices for binary prediction and multiclass prediction. The middle of each tile is the normalized count (overall percentage). The bottom number of each tile is the column percentage and the right side of each tile is the row percentage. i) decision tree, ii) logistic regression, iii) feed-forward neural net, iv) XGBoost, and v) biLSTM.*

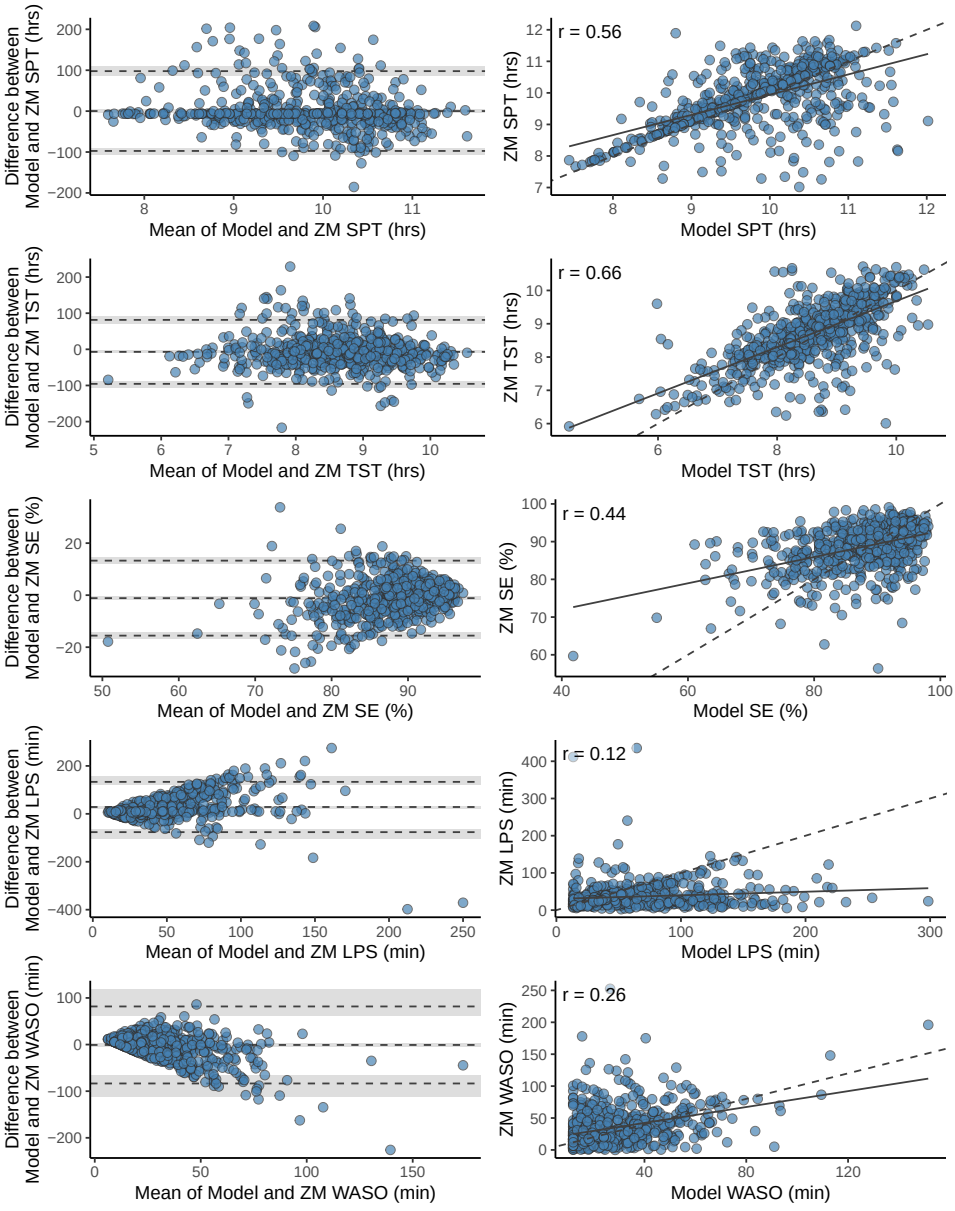

Figure 5: *Comparison of sleep quality metrics derived from the XGBoost model trained on the 5-minute smoothed ZM predictions. The left column displays Bland-Altman plots. Dashed lines represent the bias (the average difference between the two measurements) and LOA, with the 95% confidence intervals represented as the grayed areas. The right column displays scatter plots of XGBoost-derived vs ZM-derived sleep quality metrics. The dashed line represents the identity line, while the full-drawn line represents the best linear fit. Pearson's correlations are annotated in the upper left corner.*

greatest underestimation, with a bias of -36 minutes, reflecting trends observed in previous research by Winkler et al. conducted in young- middle-aged and older adults[19]. They developed an algorithm that, despite a moderate correlation (Pearson correlation coefficient = .67) between their algorithmic results and diary-recorded waking times, overestimated waking wear time by more than 30 minutes, resulting in an underestimation of in-bed time[19]. This trend was further confirmed when Inan-Eroglu et al. examined Winkler et al.'s algorithm, revealing a underestimation of 9.8 minutes in bed time compared to self-reported measures in middle-aged adults[17]. In contrast, a study by Berg et al. reported a slight underestimation of in-bed time in a sample of middle-aged and older adults. They employed a unique approach with their algorithm, which relied on quantifying the number and duration of sedentary periods to determine time in bed, and active periods (standing or stepping) to identify wake times[18]. Finally. it is important to note that predictive performance in determining in-bed time does not necessarily translate to accurate predictions of broader sleep quality metrics. The crucial task of detecting awake periods during in-bed time, a key factor in assessing further derived sleep quality metrics, is not effectively captured by in-bed time predictions alone. Furthermore, the distinction between actual sleep and time spent in bed awake, often overlooked but vital in sleep research, is critical for a comprehensive understanding of sleep quality.

To the best of our knowledge, Johansson and colleagues[20] are the only researchers who have reported epoch-to-epoch performance metrics for sleep scoring using thigh-worn accelerometers, beyond just "waking time" and "in-bed time." They achieved a mean sensitivity of 0.84, specificity of 0.55, and accuracy of 0.80, using a single-night evaluation dataset of 71 adult subjects. Despite our models achieving a sensitivity above 97%, they, like Johansson et al.'s algorithm, struggled with detecting in-bed awake epochs. This is reflected in the low specificity scores, ranging from 54.7% to 76.4%, reported in our study. The challenge of low specificity is not unique to methods using data collected from thigh-worn devices. Conley et al.'s meta-analysis[47] reported similar findings when estimating sleep using wrist-worn accelerometers among healthy adults, with a mean sensitivity, accuracy, and specificity of 0.89, 0.88, and 0.53, respectively. Furthermore, Patterson and colleagues[48] recently summarized the performance of various heuristic algorithms, machine learning, and deep learning models used to predict sleep. They found the mean sensitivity and specificity to be 93% (SD = 2.8) and 60% (SD = 11.1) respectively. These findings underscore the challenge of automating the detection of in-bed awake periods. Interestingly, despite low specificity values for most of our models and configurations, we observed an overestimation from several of our models of LPS and WASO, contrasting with most previous research[11,47]. This overestimation of wake epochs is evident from the low NPV scores, indicating that only a small proportion of the wake predictions are actually correct. This discrepancy may be driven by the SMOTE process used to balance the dataset. If the synthetic "wake" samples created by SMOTE are not representative of the true "wake" data, the models might learn to incorrectly classify certain "sleep" epochs as "wake". This could lead to an overestimation of LPS and WASO, as the models are incorrectly identifying more periods of wakefulness during the sleep period.

The use of the SMOTE technique likely improved the performance of our models by addressing the class imbalance in our data. However, this technique also introduced synthetic"wake" samples that may not be fully representative of true wake data. This could potentially lead some models to overestimate the wake class. Interestingly, the biLSTM model, which was not trained on SMOTE-processed data, was the only one to overestimate TST and SE. On the other hand, the XGBoost model, which was trained on data subjected to the SMOTE process, was able to handle the synthetic "wake" samples better than the other models, and it did not overestimate TST to the same degree. The Bland-Altman statistics for the XGBoost model trained on the 5-minute median filtered ZM predictions showed a mean difference of -7 minutes for TST and -1.1% for SE, with limits of agreement ranging from -95.5 to 81.4 minutes and from -15.6% to 13.3% respectively. This suggests that the XGBoost model was able to maintain a balance between sensitivity and specificity, and it was not overly influenced by the synthetic "wake" samples. The XGBoost model's success with the SMOTE dataset may be due to its ability to handle non-representative synthetic samples. XGBoost's gradient boosting mechanism allows it to iteratively learn from the errors of previous models, which can help it to better distinguish between true wake data and non-representative synthetic wake samples created by SMOTE. This iterative learning process could make XGBoost more robust to the inaccuracies introduced by the synthetic samples, leading to better overall performance.

Typically, sleep detection methods are applied in two contexts: either to night recordings or to 24-hour

recordings. In night recordings, it is possible to derive sleep quality metrics like SE and LPS because the SPT is already known because it is inferred from the length of the recording[47,48]. On the other hand, when sleep detection methods are applied to 24-hour recordings, most methods do not have the ability to infer the SPT without sleep diaries[49]. Consequently, these methods are unable to generate sleep quality metrics that rely on the SPT[50,51]. To overcome this limitation, we have incorporated models that can differentiate between in-bed awake time and in-bed asleep from out-bed awake time over a 24-hour recording. This approach allows our models to estimate all commonly used sleep quality metrics. Van Hees et al.[52] have proposed an algorithm to determine SPT from data collected by wrist-worn devices. This algorithm was recently validated by Plekhanova and her team[53]. By combining this algorithm with other methods, further sleep quality metrics can be inferred based on the identified SPT. Van Hees et al.[52] reported good agreements and low mean differences compared to self-report and PSG on SPT, findings later confirmed by Plekhanova and colleagues. However, they also observed poor agreement with LPS and Wake After Sleep Onset (WASO). They found low reliability with PSG, indicating difficulties in detecting wakefulness during in-bed time. These challenges parallel those we experienced in our study.

In our evaluation of sleep quality metrics, we found that LPS had the largest mean error relative to absolute time allocated to LPS. This suggests that the initial epochs of Sleep Period Time (SPT) are particularly challenging to classify correctly. This is also supported by the poor Pearson correlations between LPS derived from model predictions and the ZM. The XGBoost model, which was the best performer among all models, overestimated LPS by an average of 26.4 minutes for models trained on raw ZM predictions, 28.5 minutes for models trained on 5-minute filtered ZM predictions, and 34.5 minutes for models trained on 10-minute filtered ZM predictions. This level of discrepancy is comparable to the mean error of sleep latency of 23 minutes reported by Johansson et al.[20]. Johansson et al. suggest that the discrepancy with the gold standard is likely due to the multifaceted nature of the sleep state, which is a complex physiological process. Short awakenings or sleep episodes may not necessarily correspond to noticeable changes in thigh movement, making them difficult to detect and accurately classify. These results align with several methods for wrist-worn devices reviewed by Conley and colleagues[47]. They reported correlations between accelerometer and PSG sleep onset latency (equivalent to LPS) from 10 studies with a mean correlation of 0.2 (ranging from -0.69 to 0.69), indicating the inherent difficulty in estimating LPS using accelerometry alone.

Our study's XGBoost model demonstrated relatively narrower LOAs for TST, SE, and WASO, with ranges of -95.5 to 81.4 min, -15.6 to 13.3%, and -83.4 to 81.7 min, respectively when compared with other models such as the Van Hees algorithm[14], Oakley rsc (rescored)[11], and LSTM-50[11] evaluated in the Patterson et al. study[48]. Furthermore, comparing the LOAs between our XGBoost model and the algorithm developed for thigh-worn devices by Johansson et al. study[20], our XGBoost model showed narrower LOAs for TST , SE, LPS , and WASO, but not SPT. Generally, all methods, both from this study and from the reviewed literature, exhibit wide LOAs suggesting that there is high variability in the derived sleep quality metrics, and accelerometry cannot be used interchangeably as an alternative to the EEG-based ZM or PSG to measure sleep on an individual level. In the current study, the presence of extreme outliers seem to drive the widening of the LOAs. These findings imply that the current methods, are only reasonably accurate for assessing sleep quality metrics at a group level and caution should be exercised when applying the models and methods to individual-level sleep assessments. Therefore, further improvements and refinements are needed to enhance the validity of these models for individual sleep assessments.

In this study, we used the ZM as the reference method, rather than PSG, which is considered the gold standard for sleep measurement. This choice may contribute to discrepancies between our models and the ZM, as without a true gold standard, it's difficult to determine the source of disagreement. However, we believe that the use of ZM, which allows for multiple consecutive nights of recording in free-living[26], is valuable. This approach captures intra-individual variances in sleep, which is impractical with PSG. It also enabled us to include more nights in our study typically compared to those relying on PSG. For instance, the widely used Newcastle dataset[14] only contains data from 28 participants. However, upon examining the ZM outputs, we found that the raw predictions were not optimal for developing machine learning models due to a seemingly low signal-to-noise ratio (see Figure 3). The ZM itself mitigates this issue by applying certain filtering processes when generating sleep quality metrics. For example, epochs contributing to WASO must be in contiguous epochs of 3, and sleep only counts towards sleep quality metrics if 10 out of 12 minutes are scored as sleep. To improve the prospect of our machine learning

algorithms, we applied median filters to the ZM raw predictions. This did in fact alter the derived sleep quality metrics. Notably, the mean WASO decreased from 39 minutes in the raw predictions to 30.6 minutes in the 5-minute median filtered predictions, and further decreased to 22.3 minutes in the 10-minute filtered predictions. The application of 5-minute and 10-minute median filters also led to increases in TST, SE, and LPS. This suggests that the filters may categorize some instances of wakefulness as sleep and remove brief awakenings. Despite these changes, the overall sleep quality profile derived from the median-filtered predictions is still comparable to that from the raw predictions, justifying our approach.

The current study boasts several strengths, including the capacity to distinguish in-bed awake and asleep time from out-of-bed time, thereby allowing for the extraction of vital sleep quality metrics. Furthermore, the research benefits from evaluating multiple nights per subject, providing valuable information into intra-subject sleep variability. However, certain limitations exist. The use of ZM, which isn't recognized as a gold standard, could potentially compromise our findings' validity. Future research should consider using PSG as a reference for methods similar to ours, despite its limitations, for a more accurate comparison. Moreover, our models weren't validated using an external dataset, a process that would have showcased their broader applicability. Hence, our conclusions remain confined primarily to children.

In conclusion, our study contributes to the ongoing efforts to improve sleep estimation methods using thigh-worn accelerometers. We evaluated different machine learning and deep learning models for predicting in-bed and sleep times and their corresponding sleep quality metrics. While the sequential models generally demonstrated excellent performance in predicting in-bed time, they faced challenges in accurately distinguishing between sleep and wake epochs during in-bed time. Among all models and configurations evaluated, the XGBoost model exhibited the best performance, including epoch-to-epoch predictions and sleep quality metrics. Our research also highlighted the current limitations of sleep detection methods, such as challenges in effectively detecting wake periods during in-bed time and the need for further improvements to increase the precision of individual sleep assessments. We believe our work lays the groundwork for future research to further refine and improve the performance of these models, contributing to a more precise and accurate evaluation of sleep patterns and quality using thigh-worn accelerometers.

V. Data Availability

The data underlying this article will be shared on reasonable request to the corresponding author.

VI. Code Availability

All code used to perform the analysis and generate the figures in this paper are available in this repository. Additionally, we provide a freely available tool which allows users to employ the developed XGBoost models for estimating sleep quality metrics based on accelerometer data in this repository.

VII. Acknowledgements

This work was supported by funding from TrygFonden (grant number ID 130081 and 115606) and the European Research Council (grant number 716657).

VIII. Author Contributions

E.H.L and J.C.B designed the study. P.L.K, JS.P., S.O.S, S.R.M, and A.G. conducted the data collection. E.H.L prepared the manuscript and performed all analyses and model developments. E.H.L. and J.C.B. interpreted the results. All authors validated the methodology and approved the final manuscript.

IX. Disclosure Statement

None declared.

References

1. Ma G. Sleep, Health, and Society. *Sleep medicine clinics*. 2017;12(1). doi:10.1016/j.jsmc.2016.10.012
2. Meyer N, Harvey AG, Lockley SW, Dijk DJ. Circadian rhythms and disorders of the timing of sleep. *The Lancet*. 2022;400(10357):1061-1078. doi:10.1016/S0140-6736(22)00877-7
3. K Pavlova M, Latreille V. Sleep Disorders. *The American Journal of Medicine*. 2019;132(3):292-299. doi:10.1016/j.amjmed.2018.09.021
4. Difrancesco S, Lamers F, Riese H, et al. Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study. *Depression and Anxiety*. 2019;36(10):975-986. doi:10.1002/da.22949
5. Van De Water ATM, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography – a systematic review. *Journal of Sleep Research*. 2011;20(1pt2):183-200. doi:10.1111/j.1365-2869.2009.00814.x
6. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: A meta-analysis. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*. 2022;18(1):193-202. doi:10.5664/jcsm.9538
7. Gaiduk M, Serrano Alarcón Á, Seepold R, Martínez Madrid N. Current status and prospects of automatic sleep stages scoring: Review. *Biomedical Engineering Letters*. 2023;13(3):247-272. doi:10.1007/s13534-023-00299-3
8. Moore CM, Schmiege SJ, Matthews EE. Actigraphy and Sleep Diary Measurements in Breast Cancer Survivors: Discrepancy in Selected Sleep Parameters. *Behavioral Sleep Medicine*. 2015;13(6):472-490. doi:10.1080/15402002.2014.940108
9. Webster JB, Kripke DF, Messin S, Mullaney DJ, Wyborney G. An activity-based sleep monitor system for ambulatory use. *Sleep*. 1982;5(4):389-399. doi:10.1093/sleep/5.4.389
10. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep*. 1992;15(5):461-469. doi:10.1093/sleep/15.5.461
11. Palotti J, Mall R, Aupetit M, et al. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digital Medicine*. 2019;2(1):1-9. doi:10.1038/s41746-019-0126-9
12. Sazonov E, Sazonova N, Schuckers S, Neuman M, CHIME Study Group. Activity-based sleep-wake identification in infants. *Physiological Measurement*. 2004;25(5):1291-1304. doi:10.1088/0967-3334/25/5/018
13. Sadeh A, Sharkey KM, Carskadon MA. Activity-based sleep-wake identification: An empirical test of methodological issues. *Sleep*. 1994;17(3):201-207. doi:10.1093/sleep/17.3.201
14. Hees VT van, Sabia S, Anderson KN, et al. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. *PLOS ONE*. 2015;10(11):e0142533. doi:10.1371/journal.pone.0142533
15. Sundararajan K, Georgievska S, Lindert BHW te, et al. Sleep classification from wrist-worn accelerometer data using random forests. *Scientific Reports*. 2021;11(1):24. doi:10.1038/s41598-020-79217-x
16. Carlson JA, Tuz-Zahra F, Bellettiere J, et al. Validity of Two Awake Wear-Time Classification Algorithms for activPAL in Youth, Adults, and Older Adults. *Journal for the Measurement of Physical Behaviour*. 2021;4(2):151-162. doi:10.1123/jmpb.2020-0045
17. Inan-Eroglu E, Huang BH, Shepherd L, et al. Comparison of a Thigh-Worn Accelerometer Algorithm With Diary Estimates of Time in Bed and Time Asleep: The 1970 British Cohort Study. *Journal for the Measurement of Physical Behaviour*. 2021;4(1):60-67. doi:10.1123/jmpb.2020-0033
18. Berg JD van der, Willems PJB, Velde JHPM van der, et al. Identifying waking time in 24-h accelerometry data in adults using an automated algorithm. *Journal of Sports Sciences*. 2016;34(19):1867-1873. doi:10.1080/02640414.2016.1140908

19. Winkler EAH, Bodicoat DH, Healy GN, et al. Identifying adults' valid waking wear time by automated estimation in activPAL data collected with a 24 h wear protocol. *Physiological Measurement*. 2016;37(10):1653. doi:10.1088/0967-3334/37/10/1653

20. Johansson PJ, Crowley P, Axelsson J, et al. Development and performance of a sleep estimation algorithm using a single accelerometer placed on the thigh: An evaluation against polysomnography. *Journal of Sleep Research*. 2023;32(2):e13725. doi:10.1111/jsr.13725

21. Skotte J, Korshøj M, Kristiansen J, Hanisch C, Holtermann A. Detection of Physical Activity Types Using Triaxial Accelerometers. *Journal of Physical Activity and Health*. 2014;11(1):76-84. doi:10.1123/jpah.2011-0347

22. Arvidsson D, Fridolfsson J, Börjesson M, et al. Re-examination of accelerometer data processing and calibration for the assessment of physical activity intensity. *Scandinavian Journal of Medicine & Science in Sports*. 2019;29(10):1442-1452. doi:10.1111/sms.13470

23. Brønd JC, Grøntved A, Andersen LB, Arvidsson D, Olesen LG. Simple Method for the Objective Activity Type Assessment with Preschoolers, Children and Adolescents. *Children (Basel, Switzerland)*. 2020;7(7):72. doi:10.3390/children7070072

24. Kaplan RF, Wang Y, Loparo KA, Kelly MR, Bootzin RR. Performance evaluation of an automated single-channel sleep–wake detection algorithm. *Nature and Science of Sleep*. 2014;6:113-122. doi:10.2147/NSS.S71159

25. Wang Y, Loparo KA, Kelly MR, Kaplan RF. Evaluation of an automated single-channel sleep staging algorithm. *Nature and Science of Sleep*. 2015;7:101-111. doi:10.2147/NSS.S77888

26. Pedersen J, Rasmussen MGB, Olesen LG, Kristensen PL, Grøntved A. Self-administered electroencephalography-based sleep assessment: Compliance and perceived feasibility in children and adults. *Sleep Science and Practice*. 2021;5(1):8. doi:10.1186/s41606-021-00059-1

27. Rasmussen MGB, Pedersen J, Olesen LG, et al. Short-term efficacy of reducing screen media use on physical activity, sleep, and physiological stress in families with children aged 4–14: Study protocol for the SCREENS randomized controlled trial. *BMC Public Health*. 2020;20(1):380. doi:10.1186/s12889-020-8458-6

28. Pedersen J, Rasmussen MGB, Sørensen SO, et al. Effects of Limiting Recreational Screen Media Use on Physical Activity and Sleep in Families With Children: A Cluster Randomized Clinical Trial. *JAMA pediatrics*. 2022;176(8):741-749. doi:10.1001/jamapediatrics.2022.1519

29. Skovgaard EL, Roswall MA, Pedersen NH, Larsen KT, Grøntved A, Brønd JC. Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings. *Scientific Reports*. 2023;13(1):2496. doi:10.1038/s41598-023-29666-x

30. Walch O, Huang Y, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*. 2019;42(12):zsz180. doi:10.1093/sleep/zsz180

31. Galland BC, Taylor BJ, Elder DE, Herbison P. Normal sleep patterns in infants and children: A systematic review of observational studies. *Sleep Medicine Reviews*. 2012;16(3):213-222. doi:10.1016/j.smrv.2011.06.001

32. Pedersen MJ, Leonthin H, Mahler B, Rittig S, Jennum PJ, Kamperis K. Two nights of home polysomnography in healthy 7-14-year-old children – Feasibility and intraindividual variability. *Sleep Medicine*. 2023;101:87-92. doi:10.1016/j.sleep.2022.10.027

33. Palm L, Persson E, Elmqvist D, Blennow G. Sleep and Wakefulness in Normal Preadolescent Children. *Sleep*. 1989;12(4):299-308. doi:10.1093/sleep/12.4.299

34. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Computation*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

35. Sano A, Chen W, Lopez-Martinez D, Taylor S, Picard RW. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. *IEEE journal of biomedical and health informatics*. 2019;23(4):1607-1617. doi:10.1109/JBHI.2018.2867619

36. Chen Z, Wu M, Cui W, Liu C, Li X. An Attention Based CNN-LSTM Approach for Sleep-Wake Detection With Heterogeneous Sensors. *IEEE journal of biomedical and health informatics*. 2021;25(9):3270-3277. doi:10.1109/JBHI.2020.3006145

37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357. doi:10.1613/jair.953

38. Hvitfeldt E. *Themis: Extra Recipes Steps for Dealing with Unbalanced Data.*; 2023. https://CRAN.R-project.org/package=themis.

39. Hjorth MF, Chaput JP, Damsgaard CT, et al. Measure of sleep and physical activity by a single accelerometer: Can a waist-worn Actigraph adequately measure sleep in children? *Sleep and Biological Rhythms*. 2012;10(4):328-335. doi:10.1111/j.1479-8425.2012.00578.x

40. Kushida CA, Chang A, Gadkary C, Guilleminault C, Carrillo O, Dement WC. Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Medicine*. 2001;2(5):389-396. doi:10.1016/s1389-9457(00)00098-8

41. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical Science*. 1996;11(3):189-228. doi:10.1214/ss/1032280214

42. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2023. https://www.R-project.org/.

43. Kuhn M, Wickham H. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.*; 2020. https://www.tidymodels.org.

44. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *Journal of Open Source Software*. 2019;4(43):1686. doi:10.21105/joss.01686

45. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.

46. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019:8024-8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

47. Conley S, Knies A, Batten J, et al. Agreement between actigraphic and polysomnographic measures of sleep in adults with and without chronic conditions: A systematic review and meta-analysis. *Sleep Medicine Reviews*. 2019;46:151-160. doi:10.1016/j.smrv.2019.05.001

48. Patterson MR, Nunes AAS, Gerstel D, et al. 40 years of actigraphy in sleep medicine and current state of the art algorithms. *npj Digital Medicine*. 2023;6(1):1-7. doi:10.1038/s41746-023-00802-1

49. Girschik J, Fritschi L, Heyworth J, Waters F. Validation of self-reported sleep against actigraphy. *Journal of Epidemiology*. 2012;22(5):462-468. doi:10.2188/jea.je20120012

50. Doherty A, Jackson D, Hammerla N, et al. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE*. 2017;12(2):e0169649. doi:10.1371/journal.pone.0169649

51. Anderson KN, Catt M, Collerton J, et al. Assessment of sleep and circadian rhythm disorders in the very old: The Newcastle 85+ Cohort Study. *Age and Ageing*. 2014;43(1):57-63. doi:10.1093/ageing/aft153

52. Hees VT van, Sabia S, Jones SE, et al. Estimating sleep parameters using an accelerometer without sleep diary. *Scientific Reports*. 2018;8(1):12975. doi:10.1038/s41598-018-31266-z

53. Plekhanova T, Rowlands AV, Davies MJ, Hall AP, Yates T, Edwardson CL. Validation of an automated sleep detection algorithm using data from multiple accelerometer brands. *Journal of Sleep Research*. 2023;32(3):e13760. doi:10.1111/jsr.13760