COMPUTATIONAL SLEEP SCIENCE:

MACHINE LEARNING FOR THE DETECTION, DIAGNOSIS, AND TREATMENT OF SLEEP PROBLEMS FROM WEARABLE DEVICE DATA

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Aarti Sathyanarayana

IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jaideep Srivastava

December 2017

Preface

This thesis is motivated by the rapid increase in global life expectancy without the respective improvements in quality of life. I propose several novel machine learning and data mining methodologies for approaching a paramount component of quality of life, the translational science field of sleep research. Inadequate sleep negatively affects both mental and physical well-being, and exacerbates many non-communicable health problems such as diabetes, depression, cancer and obesity. Taking advantage of the ubiquitous adoption of wearable devices, I create algorithmic solutions to analyse sensor data. The goal is to improve the quality of life of wearable device users, as well as provide clinical insights and tools for sleep researchers and care-providers.

- Chapter 1 is the introduction. This section substantiates the timely relevance of sleep research for today's society, and its contribution towards improved global health. It covers the history of sleep science technology and identifies core computing challenges in the field. The scope of the thesis is established and an approach is articulated. Useful definitions, sleep domain terminology, and some pre-processing steps are defined. Lastly, an outline for the remainder of the thesis is included.

- Chapter 2 dives into my proposed methodology for widespread screening of sleep disorders. It surveys results from the application of several statistical and data mining methods. It also introduces my novel deep learning architecture optimized for the unique dimensionality and

nature of wearable device data.

- Chapter 3 focuses on the diagnosis stage of the sleep science process. I introduce a human activity recognition algorithm called RAHAR, Robust Automated Human Activity Recognition. This algorithm is unique in a number of ways, including its objective of annotating a behavioural time series with exertion levels rather than activity type.

- Chapter 4 focuses on the last step of the sleep science process, therapy. I define a pipeline to identify *behavioural recipes*. These *recipes* are the target behaviour that a user should complete in order to have good quality sleep. This work provides the foundation for building out a dynamic real-time recommender system for wearable device users, or a clinically administered cognitive behavioural therapy program.

- Chapter 5 summarizes the impact of this body of work, and takes a look into next steps. This chapter concludes my thesis.

Acknowledgements

Firstly, I would like to thank Dr. Jaideep Srivastava, my advisor and friend, for his limitless support over the past 5 years. Thank you for pushing me to be the best that I can be. You have an ability to see the big picture and the minute details simultaneously, to formulate and communicate a vision across differing disciplines, and to show up everyday with an infectious smile on your face. Your humble disposition, your eager attitude, and your thirst for knowledge has inspired me both professionally and personally. Thank you for always believing in me.

I would also like to thank my committee members, whose guidance has helped form this body of work. I am so grateful for their time. Dr. Vipin Kumar and Dr. Gyorgy Simon have made significant contributions to my academic growth for the last 5 years. Thank you both for teaching me, for guiding me, and for helping me rediscover my path when I felt overwhelmed. I have learnt so much from you both. Thank you to Dr. Louis Kazaglis for sharing your expert knowledge in the area of sleep science. Your commitment to the field is what brings new research ideas to the real world's clinics. Thank you also to Dr. Abhishek Chandra. I greatly appreciate your support and guidance in writing this thesis, and for supporting my nomination for the University of Minnesota's Doctoral Dissertation Fellowship.

Thank you to the University of Minnesota, particularly the Department of Computer Science. You have been my home base for many years. Thanks to Sara Howard. I wouldn't be making it to

graduation, if it wasn't for her diligence. Thank you to Phillip Barry for his work behind the scenes to assign teaching opportunities. And thank you to each Professor that I have TA'd for: Dr. John Carlis, Dr. Steven Jensen, Dr. Shashi Shekhar, and Dr. Carl Sturtivant. You have taught me to teach, which is invaluable.

Outside of the University of Minnesota, I have been fortunate to receive advice, support, and friendship from many. To Dr. Luis Fernandez Luque, thank you for solidifying my interest in human health and wellness. You are a never-ending resource of information, and your passion is inspiring. Your encouragement has pushed me to work harder than ever in the past two years and I sincerely look forward to years of continued collaborations! I would also like to thank Meghna Singh. You are truly great at what you do. Thank you for all the engineering time you have dedicated behind the scenes. Dr. Ahmed Elmagarmid, thank you for believing in me. Working at the Qatar Computing Research Institute was an experience of a lifetime, and I will not forget the friends and colleagues I met while living in a small country across the world. You have always pushed me to dream big and I won't stop now. To Dr. Jyotishman Pathak, it was a pleasure to work with you at Mayo Clinic. Your continued support over the years has been deeply appreciated. To Dr. Prasanna Desikan, your ability to thrive in interdisciplinary roles has been a source of inspiration for me. I enjoyed our time at Allina Health together. To TC Tong, thank you for trusting in my skills and for showing me a new side of healthcare research. Thank you to my co-authors, Dr. Ferda Ofli from Qatar Computing Research Institute and Dr.Shafiq Joty from Nanyang Technological University for their hard work and of course hard criticism! I would also like to thank Dr. Shahrad Taheri and Dr. Teresa Arora from Weill Cornell Medical College for contributing their expertise in the area of sleep science, and for providing the clinical trial data used in my research. I would also like to thank Brent Ledvina from Apple, for teaching by example. I can only hope to be as great of a manager as you someday.

I would also like to thank and dedicate this doctoral dissertation to my parents, Babu & Jyoti Sathyanarayana, and brother, Nikhil. You give me courage and significance. I live my life everyday

in respect of you. Thank you for raising me, encouraging me, and loving me. We have been through so much together, but we have come out on top. Dad, this Ph.D has been as much your dream as it has been mine. Thank you for being my life-long tutor. Your job is not yet done! Mum, you've taught me to think outside the box, to push the boundaries, and to go after what I want. Thank you for shaping me into who I am today. To my little brother Nikhil, you are irreplaceable. Thank you for your unconditional support. It means the world.

To Dr. Borislav Alexandrov, my love. You haven't just supported my dreams, you've been a helping hand every step of the way. Thank you for all that you are.

Lastly, thank you to all others who have contributed directly and indirectly towards my education, academic pursuits, and general happiness. Thank you to each of my friends, lab-mates and colleagues. This is but a milestone on a long journey. As the good Bertrand Russell said, "The good life is one inspired by love and guided by knowledge".

Dedication

To Babu, Jyoti & Nikhil Sathyanarayana

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There has been a recent epidemiological transition in the leading causes of death, from acute infectious diseases to chronic, non-communicable diseases [3] [4]. These diseases currently impose the most significant burden on national and global healthcare [3] [5] [6], with an estimated cumulative output loss over the next two decades, as high as $47 trillion [5]. This vast sum is due in part to the fact that the 20th century saw a rapid increase in global life expectancy, without the respective improvements in quality of life [3] [4]. The population ageing trend has continued, even accelerated, into the 21st century. Though the extension of mortality is indeed a huge achievement in human health, it introduces several practical and economic challenges, for individuals as well as society.

Among the practical challenges of an increased lifespan, the most critical is the quality of latter life. The benefits of a longer lifespan cannot be realised if it is a life of illness and disability [3, 7, 8]. Moreover, non-communicable diseases often require expensive and extensive care. They are slow in progression, and without proper continuous care can lead to many forms of disability. A society with a growing elderly or disabled population will find itself tempered in the productivity of its labour force, and the economic expansion of its markets [5].

It is no secret that health care is expensive and costs are increasing [5] [9]. At an individual level, the financial and opportunity costs of healthcare can be debilitating. This often results in inadequate treatment and care for those belonging to particular socio-economic levels [3, 7]. This is especially true, because non-communicable diseases are often triggered into exacerbation by external factors such as environment and lifestyle. Studies in health economics explore economically viable treatment options stratified by efficacy and cost, but running clinical trials can be exorbitant, and the results are limited by short and artificially simulated periods of evaluation.

There are two general approaches that can reduce the impact of non-communicable diseases: (1) delaying the severity of disability, thus allowing the ageing population to remain independent for as long as possible, and (2) delaying the onset of the disease through preventative care, specifically in children and adolescents.

Both delaying the severity of disability and delaying disease onset require early detection and diagnosis [3]. Obesity, diabetes, chronic respiratory diseases, and even some types of cancer can be avoided through appropriate personalised and preventative care [3] [10]. Moreover, recent evidence indicates that it may even be possible to delay the onset of serious degenerative diseases such as Alzheimer's and Parkinson's [11] [12].

The *All of Us* research program, originally named *The Precision Medicine Initiative*, is a $215 million research investment initiated by President Obama in his 2015 State of the Union. The program was formed to amass genetic, environmental, and lifestyle data from a large national research cohort with the objective of personalising healthcare. An important component of the National Institute of Health's implementation of this effort is to collect data from biometric and physiological sensors, such as wearable devices and mobile phones [13]. Sensor data is crucial to the personalisation of care, because it provides ambulatory and real-time monitoring. This is especially relevant for the treatment of non-communicable diseases, where a longitudinal understanding of behaviour can greatly improve preventative care [14]. While the *All of Us* research program may be the largest,

it is only one of many such initiatives. The *Kavli HUMAN Project* collects data from 2,500 New York households over 20 years [15]. The *a 360° Quantified Self Program* was used to collect (as well as influence) nutrition, physical activity, parental involvement, biometric information, and social media behaviour of a group of Qatari adolescents continuously for a 6 week summer camp and then intermittently for years beyond [16].

The rising popularity of wearable devices makes them the perfect tool for longitudinal and ubiquitous sensing. In 2014, 1 in 6 consumers owned a wearable device [17]. Today the wearables market is projected to exceed $4 billion by the end of 2017 [18]. These cheap and ubiquitous sensors feel intrinsic and unobtrusive to the user and can thus reveal new insights into human behaviour in an unprecedented manner. However, continuous sensing generates large amounts of big data that requires advanced computational analysis [9]. The predictive power provided by machine learning algorithms can aid translation of this enigmatic behavioural monitoring into medical knowledge discovery.

1.1 The Effect of Sleep on Quality of Life

The importance of a good night's sleep is paramount to quality of life. Insufficient sleep can impede physical, emotional, and mental well-being [19, 20], and lead to a variety of health problems such as insulin resistance [21], cardiovascular disease [22], mood disorders (e.g., depression or anxiety) [23], and decreased cognitive function for memory and judgement [24]. This is even more apparent in shift workers who suffer from disruption to their circadian rhythm [25]. In professions such as operating construction cranes or driving public transportation, it can be life threatening. In the US alone, an estimated 810,000 sleep-related collisions in the year 2000, resulted in 1,400 fatalities and cost $16 billion [26].

Moreover, the rapid pace of modern existence has resulted in an increasing prevalence of poor sleep quality and boosted interest in studying sleep behaviours and their contributing factors [27].

Increased knowledge of the importance of sleep is fostering the inclusion of sleep education as part of patient education for diseases such as diabetes, strokes, or atrial fibrillation [28]. In addition, healthcare providers are developing applications for sleep coaching patients with sleep disorders [29, 30]. or with other diseases such as cancer [31].

The concept of focusing on sleep as a tool to improve quality of life, has recently gained a lot of attention. Ariana Huffington famously stepped down as Editor in Chief at *The Huffington Post* in 2016 to start a health and wellness startup called *Thrive Global*. She published *The Sleep Revolution: Transforming Your Life, One Night at a Time*, a book dedicated to the pertinence of sleep and its importance to achieving overall happiness and success [32]. She was not the only one to recognise the timely significance of sleep.

Large technology companies such as Apple [33], Google [34], Samsung [35], Nokia [36], Phillips [37], and Fitbit [38] are investing in sleep technology and physical behaviour tracking. There have also been a large number of startups with a focus on sleep tracking, such as SleepScore Labs and SleepAxis. In 1990, the National Sleep Foundation was formed [39]. This foundation is a non-profit organization advocating for society's better understanding of sleep. They formed *Sleep Health*, a peer-reviewed journal, and *Sleep.org*, a public resource for information about sleep. Sleep technology is a rapidly progressing area, and there is an urgent need for computational innovation.

1.2 The Evolution of Sleep Science

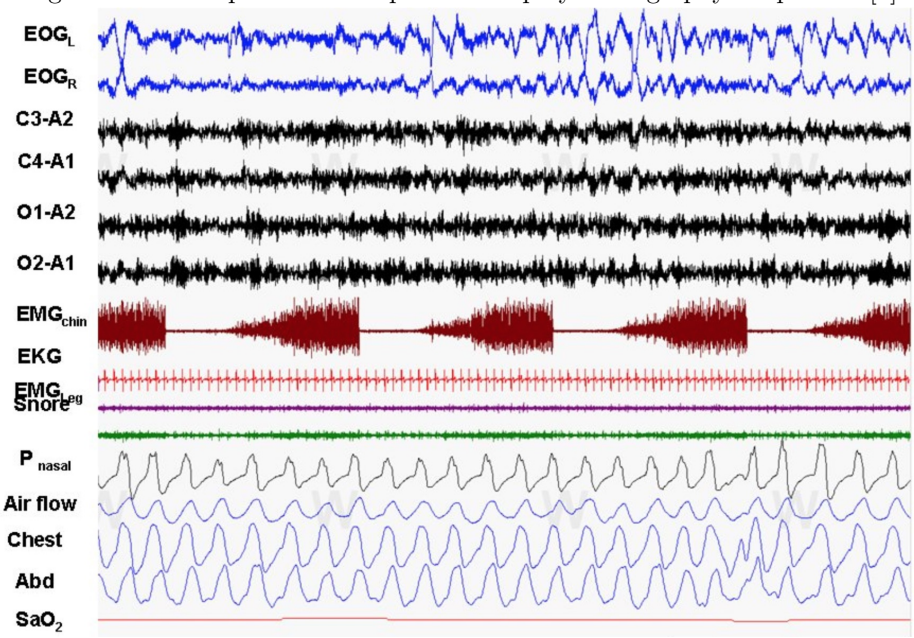
For centuries, researchers have been exploring the elements of sleep. In Bioelectromagnetics, published in 1791, Luigi Galvani explained how electric currents from neurons passed signals to active muscles in frogs. His work relating electricity to the nervous system, eventually led to Hans Bergers 1929 report of the first human electroencephalograms (EEGs) for brain study. A decade later, Alfred L. Loomis published a study describing different EEG patterns during sleep, which became the basis for defining sleep stages. Gradually, new signals evolved to characterize other sleep aspects, such as

Rapid Eye Movement (REM).

The next three decades saw an explosion of sleep knowledge, which led to a new era in sleep science and its clinical aspects. In 1960, G.W. Vogel observed that narcolepsy patients have different EEG patterns during sleep as compared to healthy subjects. In 1968, Rechtschaffen and Kales created the first definition of sleep stages [40]. By the mid-1900s, sleep laboratories began to appear that were dedicated to studying sleep-related disorders such as restless legs syndrome, sleep-obstructive apnea, narcolepsy, and insomnia. These labs resulted in more systematic sleep studies and the development of polysomnography (PSG), a sleep-study methodology that incorporates various sensors to capture brain activity, leg movement, heart rate, oxygen saturation, breathing frequency, and snoring. [41] Although it is a major diagnostic tool that is considered the gold standard for sleep research, PSG is performed only in sleep laboratories in hospitals. During a PSG test, the patient spends at least one night in the unnatural environment of a sleep laboratory monitored by a clinician and attached to multiple sensors. On one hand, PSG is a high-fidelity test; on the other hand, it is not scalable, too expensive, does not consider physical activity (which is tightly linked to sleep quality), and does not provide a holistic window into a patients behaviours in their natural everyday environment. As a result, PSG cannot keep pace with the growing number of people suffering from sleep disorders, and large segments of the population still suffer from inadequate sleep.

More recently, Home Sleep Tests (HSTs) have been proven as an effective alternative to PSG [42]. HSTs still require an in clinic visit with a specialist, but rather than spending the night at a clinic, patients take home a small device. Typically, the device consists of a small nasal cannula to measure airflow, a chest belt to measure respiration, and a finger clip to measure the blood oxygen saturation. Since HSTs use far fewer sensors than PSG, they are limited to diagnosing only sleep apnea. Whilst they are a much more affordable alternative, the lack of clinical monitoring or computational automation, means that HSTs make no distinction between awake time and sleep time. This leads to sleep apnea being under-diagnosed or mistakenly being considered as lower

Figure 1.1: Example sensor output from a polysomnography sleep clinic [1]



severity, particularly in patients that are slow to fall asleep.

Wearable devices provide the first hope of solving these problems because they can be used to study sleep and physical activity for longer periods of time and outside the laboratory or hospital. Because a wearable monitors the patients body continuously, collected signals can be more effectively used to determine sleep quality and screen for sleep disorders.

Actigraphy devices, such as the GT3X from ActiGraph, are clinical-grade wearables approved for sleep studies that use inertial sensors to collect physical activity and sleep data. As of January 2017, actigraphy was being used in more than 100 clinical trials registered in the US clinical trials database (www.clinicaltrials.gov). In many applications, actigraphy can be a cheaper and simpler alternative to PSG. In fact, actigraphy devices are already advancing sleep science in areas such as obesity, diabetes, cancer, mental health, and public health.

Figure 1.2: ResMed Home Sleep Test Device [2]

Sleep analytics is no longer restricted to researchers and clinicians, as the pervasiveness of mobile health and wellness applications attests. Millions of people use affordable wearables to track their physical activity and sleep, while companion applications integrate tracked data into mobile health repositories, such as Apples HealthKit. These repositories enable integration with other medical devices, such as heart rate monitors, and even with electronic health records (EHRs). The Apple Watch has an instrument suite to collect data on various bodily functions, including triaxial movement and heart rate. The watch seamlessly syncs with iOS devices, such as iPhones and iPads. With appropriate permissions, Apples HealthKit API allows queries to the collected data. Other smart-device manufacturers, such as Samsung and Microsoft, have created wearable platforms based on their respective operating systems. Specialized wearable device companies like Fitbit have developed their own platforms. Along with the myriad of fitness trackers and smart watches that can monitor sleep, are context sensors such as SleepScore's Max device (http://www.sleepscore.com/sleepscore-max-sleep-tracker/), Nokia's newly acquired Withings device *Aura* (www.withings.com/ca/en/products/aura/sleep-sensor-accessory), or Apple's acquisition of *Beddit* (www.beddit.com). Both include a sensing mat for sleep monitoring, and Aura integrates with the Nest Internet of Things (IoT) platform (www.nest.com) to automatically adjust room

temperature for optimal sleep quality.

The importance of data for health research has become mainstream in recent years, and sleep science is no exception. The National Institutes of Health (NIH) created the National Sleep Research Resource [43, 44], a portal aimed at integrating heterogeneous data sources for clinical sleep research (www.sleepdata.org). The portal, part of the well-known Big Data to Knowledge (BD2K) initiative, contains a wide variety of datasets for sleep research.

1.3 The Future of Sleep Science

Sleep research is a translational science field. Even the contributing medical disciplines are vast, stemming from endocrinology and pulmonology to public health. Moreover, the burst of consumer interest in the *quantified self* movement, shows a demand for a fundamental transition to patient-centric custom healthcare. The individual no longer relies, or wants to rely, on a clinic visit to improve their health. Physical fitness and nutrition tracking are two components of a person's health that personal digital devices have revolutionized. The future of sleep science looks to follow a similar trajectory. At home sleep tracking devices are quickly entering the public's vernacular. These devices empower users to track their sleep patterns and optimise their behaviours to improve their sleep quality. For the average user, the level of intervention provided by these devices is sufficient. For users with more serious sleep disorders, these devices could identify those issues, triggering a clinic visit. Furthermore, the information tracked by these devices could equip clinicians with additional lifestyle information on their patients.

Whilst machine learning algorithms applied to wearable device data have been used to study patterns and characteristics [45, 46, 47, 48], the accelerating accumulation of this big data demands improved computational technologies. High-fidelity, automated, robust, and interpretable models are vital to its success. However, big data management is not the only area of computer science that is required to advance sleep medicine. The following section addresses some of the key computing

challenges for sleep science.

1.4 Computing Challenges

The area of sleep science has many dimensions, each of which could benefit from a computer science sub-speciality.

1.4.1 Electronics and Embedded Systems

Ubiquitous sensing and the *Internet of Things* movement call on the design and development of high-fidelity sensors and actuators to capture contextual information about sleep. Creating desirable and effective wearable devices has become a central focus of many of the large technology companies of today. Increasing battery life, improving comfort, and tracking more information, are a few of the biggest hurdles to market penetration.

1.4.2 Human Computer Interaction

The development of high-fidelity sensors requires a partnership with human computer interaction (HCI) research. HCI attempts to improve the usability of devices by understanding human behaviour. For example, nutrition tracking applications such as MyFitnessPal or LoseIt suffer from churning users because manually entering everything you eat, can be tiresome. Likewise, it is unnatural for users to identify the exact moment they attempt to go to sleep or wake up. Many users watch TV in bed, or check emails in the morning. For sleep research, wearable devices must be natural and comfortable. Those that require frequent charging can result in data lapses and inconsistent use.

1.4.3 Data Integration and Interoperability

As consumer health, wellness, and fitness data is joined with electronic medical records, integration is becoming more problematic. The manner in which the unstructured data is stored within a database needs to be easily accessible and interpretable to clinicians and researchers alike. The demand to fuse information from multiple wearable sensors, such as an accelerometers and heart-rate monitors, requires new algorithms. Research is still far from producing fully integrated smart analytics for heterogeneous sleep-related data sources, but computer science can further reduce the gap.

1.4.4 Machine Learning and Artificial Intelligence

The current pervasive adoption of wearable devices provides a unique opportunity for data analytics. Wearable devices monitor a user for an extended period of time and generate a large amount of data. High-fidelity, robust, automated, and easily interpretable models are a must.

The current sleep analysis processes for actigraphy data include manual components that are unable to scale, creating a bottleneck for sleep research. Moreover, this leaves the data interpretation prone to human error. Automation allows for immediate analysis of large-scale clinical trials, and provides a platform for affordable widespread population screening of sleep disorders. Particularly for population screening, it is critical that the methods and knowledge extracted are generalizable across populations (and devices) and thus robust to noise and variance amongst sub-populations. For example, teenagers in Qatar follow very different sleep patterns to teenagers in the UK. For computational models to effectively extract actionable medical knowledge, it is critical that the models are interpretable by health professionals. Without actionable insights, society cannot benefit from sleep research discoveries. Furthermore, automation improves consistency across individuals and datasets.

Data visualization is a further sub-speciality that is needed for the interpretability of the data. Presenting real-time insights to wearable device users can influence their behaviour and guide them

towards a healthier lifestyle. Sharing information clearly with clinical professionals can allow doctors to quickly identify poor habits or identify more serious medical disorders, such as sleep apnea.

1.5 Approach

In this thesis, we focus on the machine learning approach to (*i*) screen, (*ii*) diagnose and (*iii*) provide therapy, for patients with sleep problems.

1.5.1 Current Sleep Science Process

Screening

Widespread sleep disorder screening has not yet been feasible, due to a lack of automation. Therefore current sleep science tools do not differentiate between screening and diagnosis. Sleep problems are self-reported by patients, appointments with a sleep specialist require a referral from a primary care physician, and any type of intervention or analysis, requires active involvement from a specialist. As a result, sleep clinics are overbooked and there is a long waitlist causing a bottleneck in patients receiving a formal diagnosis. Moreover, the costs incurred from a sleep clinic visit can be very expensive.

Diagnosis

The gold standard for clinical sleep diagnosis is Polysomnography (PSG). As previously mentioned, PSG is a diagnostic tool which incorporates multiple channels from various sensors. It must be performed at a sleep laboratory in a hospital, or at a clinic, and a patient must spend at least one night monitored continuously by the clinician and attached to multiple sensors to track: electroencephalography, electrocardiography, leg movement, oxygen saturation, heart rate, breathing frequency, microphones for snoring, etc. [45].

PSG is a challenging and expensive solution. Although it provides a high-fidelity test, it cannot scale to tackle the growing prevalence of sleep disorders. Furthermore, PSG cannot consider other behaviours, such as physical activity, that are highly linked to sleep quality. This is a major limitation of PSG on providing insights into the behaviours of patients within their natural environment and daily routines. As a result of these shortcomings, sleep disorders often continue without diagnosis or therapy, leading to large segments of the population suffering from inadequate sleep. It is from this context that sleep science has been driven towards the use of wearable devices.

Actigraphy is another diagnostic technique. It is specifically useful for the extended study of sleep and physical activity patterns using clinical-grade wearable devices. The devices utilize inertial sensors such as accelerometers or inclinometers, to track behaviour and activity levels. An example is the GT3X from ActiGraph Corp, which is a medical actigraph device clinically validated for sleep studies.

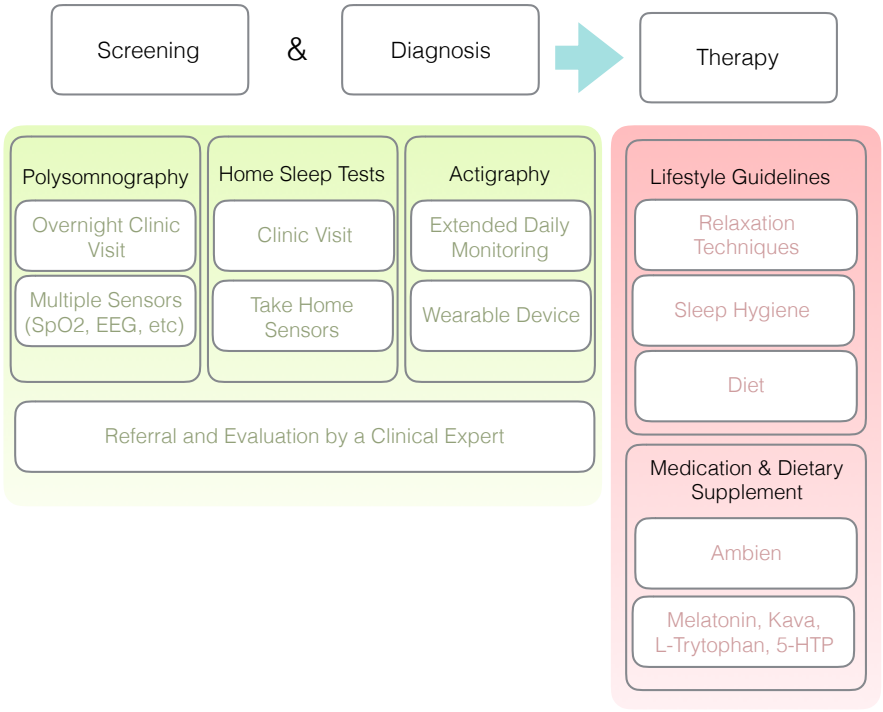
Actigraphy, in many applications, can be a cheaper and simpler alternative to PSG, and is facilitating the advancement of sleep science in areas such as, obesity, diabetes, cancer, mental health and public health. Currently, actigraphy is used in over 100 clinical trials registered in the US Clinical Trials Database (clinicaltrials.gov). Data analytics has been widely used to study sleep patterns from actigraphy data [9-12], but primarily to study different sleep patterns and characteristics [46, 47, 48, 49].

Therapy

As for the available therapies for sleep conditions, there are a variety of approaches.

Serious sleep conditions such as insomnia or sleep apnea, can require surgery, creating an oral appliance to be worn overnight, cognitive behavioural therapy or some other clinical care. For more minor sleep conditions, techniques fall into 2 major categories, (*i*) medication & dietary supplements and, (*ii*) lifestyle changes.

Figure 1.3: The Current Sleep Science Process



Ambien, also known an Zolpidem, is a sedative and hypnotic medication that holds 90% of the $2.1 billion prescription sleep medication market [50]. Ambien is a short-term treatment of sleep disorders, so it does not have prolonged improvements on sleep quality after the use of the medication is discontinued. Moreover Ambien has a short half-life and is thus not effective for many patients, who still do not sleep well through out the night. It also requires a prescription, and thus a clinic visit with a sleep specialist, and a referral from a primary care physician.

Melatonin, is a hormone that regulates the body's sleep cycle and is available as a dietary supplement from any drug/pharmacy store. While it is proven safe for short term use, the long-term effects have not been studied exhaustively. Other supplements such as valerin, kava, chamomile,

5-HTP, L-tryptophan, can have serious side effects. Kava has been linked to severe liver failure [51] and L-tryptophan has been linked to eosinophilia-myalgia syndrome (EMS) [52].

In addition to tablets, there are generalized guidelines that are suggested to improve sleep quality via lifestyle changes. Relaxation techniques such as yoga, meditation, hynotherapy, massage therapy, acupuncture and aromatherapy are all thought to improve the quality of sleep. Improved sleep hygiene (i.e no use of blue light devices such as a phone/TV/computer before bed, sleeping in an entirely dark environment, only engaging in mild activity before bed, following a regular sleep schedule including on weekends etc.) also help the quality of sleep. There are also dietary guidelines such as staying away from caffeine, alcohol and heavy meals. These guidelines are made for the overall population and are not personalized to an individual. What is more, these guidelines are often not based on strong evidence obtained through transparent research or registered clinical trials.

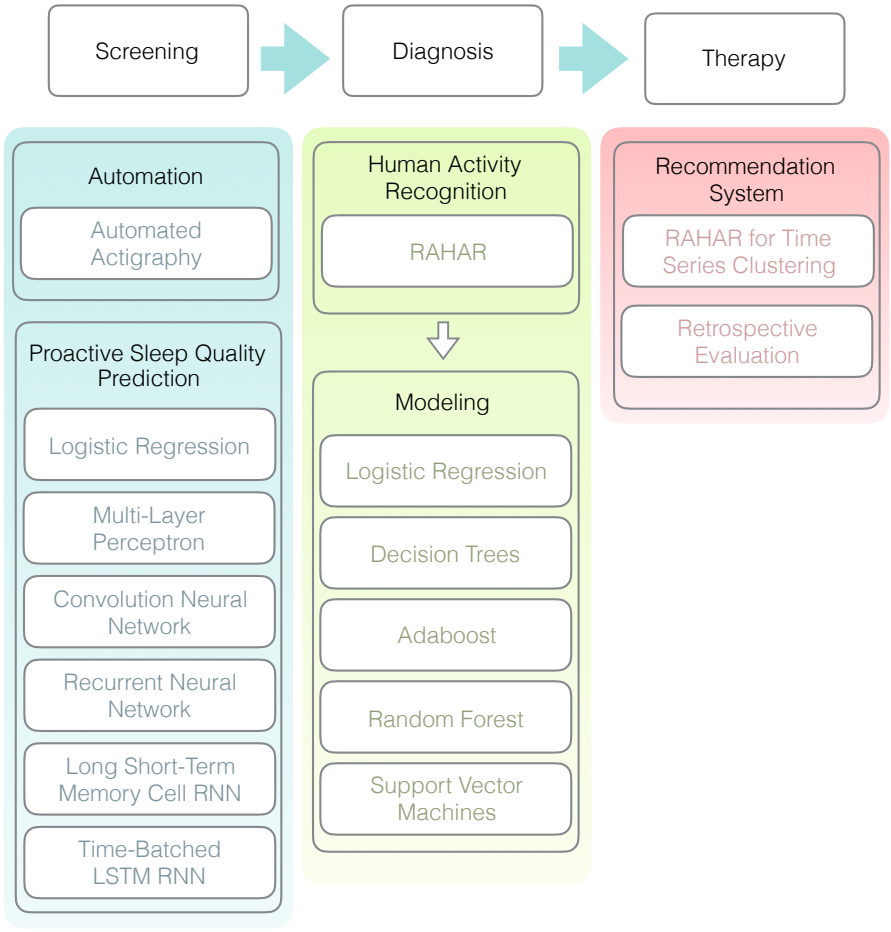
A special form of cognitive behavioural therapy for insomnia (CBT-I) is a clinical technique used specifically for treating insomnia patients. It has been proven effective in coaching individuals towards an improved lifestyle for better sleep. A therapist or sleep expert works with a patient over a number of weeks to improve their habits. However, gaining access to CBT cohorts can often be inaccessible or a bottleneck to treatment. Not to mention, CBT requires extensive involvement from a professional.

1.5.2 Proposed Revised Sleep Science Process

Figure 1.4 shows a modified version of figure 1.3 with computer science data-driven approaches to the current sleep process pipeline. These methods take advantage of (*i*) the big data available in sleep science clinical trials and wearable devices, (*ii*) the cheap relative cost of data storage and computation, and (*iii*) the public acceptance of advanced "black box" algorithms.

Not included in figure 1.4 is the concept of automated actigraphy. Actigraphy currently requires the presence of a clinical professional to manually annotate and review the output. This is

Figure 1.4: Revised Sleep Science Process



a huge bottleneck for evaluating clinical trial research, causes added expense and overhead time for clinic visits, and eliminates any possibility of widespread sleep problem screening. Thus automated actigraphy is a fundamental necessity to the sleep research community and allows for sleep quality evaluation on an individual's sleep period for wearable devices that are worn to bed. Section 2.2 dives into the details of this tool.

Screening

While automated actigraphy is a critical tool to sleep researchers, wearables often need charging overnight (e.g. Apple Watch). Even more importantly, actigraphy is a retrospective evaluation of activity, i.e. it allows for sleep quality evaluation only after the fact. I propose using data mining and machine learning to provide a prediction framework, so that sleep quality can be determined beforehand. This opens the door to the prevention of poor sleep, rather than the deduction. Classic statistical modelling methods such as linear (numerical) or logistic (classification) regression can be used for prediction. Data mining methods such as decision trees, random forest and support vector machines can also be used.

Deep learning has been a hugely popular methodology in the computer science industry over the last few years. The mathematical nature of deep learning makes it ideal for high-fidelity screening of sleep disorders. Models such as the universal approximator, a multi-layer perceptron can be used, as well as convolution neural networks, recurrent neural networks, and long short-term memory cells. In addition, I have designed a new deep learning architecture that is designed specifically with the nature of wearable actigraphy data in mind.

Diagnosis

Although deep learning has high predictive value, the justification of its prediction is not transparent. Deep learning has no beta coefficient equivalent. On the contrary, traditional statistical and data mining methods do provide insight into the prediction. The fact that traditional methods suffer in accuracy relative to deep learning, is due to the complex shape and nature of data. However, this complexity can be managed by conducting feature construction on the raw wearable device data. These modified learning representations of the raw data can improve downstream analysis and lead to improved results from traditional data analysis methods, and improved transparency relative to deep learning algorithms.

Human Activity Recognition (HAR) algorithms aim to classify and label human behaviour, from data captured by pervasive sensors, such as cameras or wearable devices. These algorithms are a powerful tool when consistent and continuous patient monitoring results in large longitudinal data collection. I propose utilizing a human activity recognition on the raw accelerometer output to create a coarser feature space that significantly improves the capabilities of traditional data mining methods.

Therapy

Lastly, I propose tools that can be used within a real-time recommendation system. This recommendations system could be used as a sleep coach within either (*i*) an application for consumer health fanatics to self-improve their sleep quality, or (*ii*) an automated alternative, or assistant, to cognitive behavioural therapy interventions. All the previously mentioned techniques provide insights into sleep quality before it occurs, but immediately before it occurs, allowing for no behavioural adjustment. By evaluating an individuals daily activities, using human activity recognition, we can identify *behavioural recipes* that lead to good or poor quality sleep. These *recipes* can then be used as target behaviour for users throughout the day, to obtain a good night sleep. In order to identify the *behavioural recipes*, I propose using clustering on a time series constructed from a human activity recognition alphabet. See chapter 4 for more details.

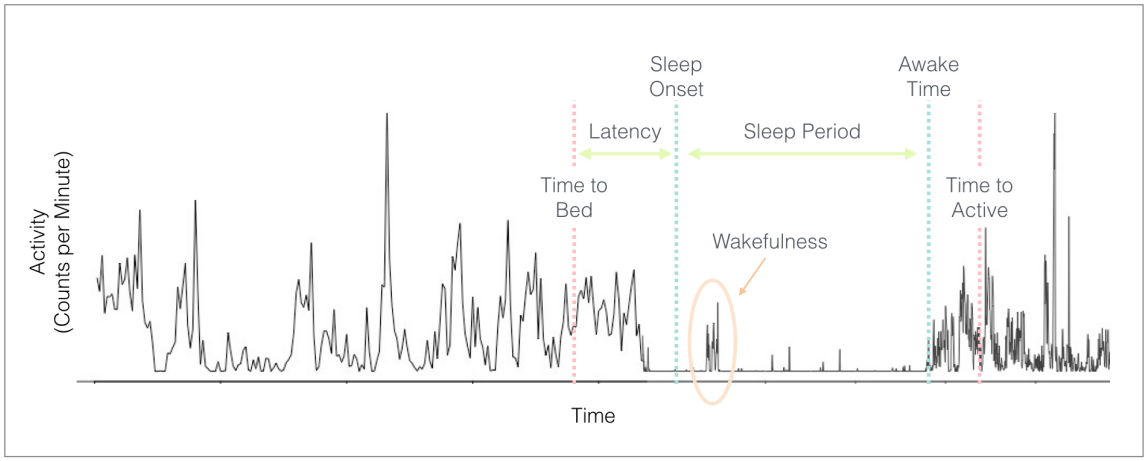
1.6 Sleep Science Metrics and Terminology

Throughout the remaining chapters, domain specific terminology from the sleep science community is used. To clarify this vocabulary for the reader, these terms are defined in this section. Figure 2.7 contains an annotated version of a sample activity time series collected from a wearable device.

Actigraphy: As aforementioned, actigraphy is the study of sleep-related behaviour measured via a clinically-validated wearable device. It is used as a tool to gain insights into an individual's

physical activity and its effect of their sleep.

Figure 1.5: Sample Accelerometer Output with Sleep Definitions Annotated



Actigraph: An actigraph is a clinically-validated wearable device. These devices contain an accelerometer, an inclinometer, and often other sensors such as a heart rate sensor, luminosity sensor, etc. The distinction between an ordinary consumer wearable and an actigraph, is in the validation. These devices are used as a clinical diagnostic tool, and for data collection in clinical trials. They are often tied to a software suite of tools for clinicians to conduct analysis. Figure 1.6 is an image of an actigraph from the company Actigraph. The device is called an Actigraph GT3X+, and contains an accelerometer, inclinometer and luminosity sensor to measure ambient light.

Activity Time Series: The output from a wearable device includes the accelerometer movement, inclinometer movement, and any other sensor information the device contains, tracked over a continuous period of time. From the accelerometer output, counts are computed to represent the frequency and intensity of the raw acceleration in epochs. Figure 2.7 graphs an activity time series, with time on the x-axis and counts per minute on the y-axis. Higher counts represent more intense movement.

Figure 1.6: The Actigraph GT3X+

Time to Bed: The time to bed is the time that an individual attempts to fall asleep. In traditional sleep science, this value is self-reported, and is thus often unreliable and inaccurate. In section 2.2, an automated actigraphy approach to determining the time to bed is described.

Sleep Onset Time: As opposed to *time to bed*, the sleep onset time is not self reported. It it the point in time where an individual actually falls asleep. In traditional actigraphy, the sleep onset time is calculated as the first minute following the self-reported time to bed, that precedes 15 continuous minutes of sleep, i.e. minimal movement tracked by the accelerometer. It also marks the start of the *sleep period*.

Latency: The difference in time between *time to bed*, and the *sleep onset time*, is referred to as the latency. In other words, latency is the amount of time it takes for an individual to fall asleep after they intend to. Whilst sleep quality is objectively measured through a variety of metrics, latency is often a good measure of the perceived sleep quality.

Awake Time: Once an individual is asleep, the moment in time that they awaken is referred

to as the *awake time*. In traditional actigraphy, this value is self-reported.

Sleep Period: The sleep period is the specific period in time that is indexed by the *sleep onset time* and the time that an individual awakens.

Time to Active: Whilst the *awake time* is self-reported, *time to active* is computed from the actigraphy signal. More details on how this value is computed, are included in section 2.2.

Wakefulness and Wake After Sleep Onset: Accelerometers within actigraph devices, are very sensitive to movement. This is so that sedentary behaviour can be differentiated from actual sleep. However, when an individual is asleep, there is often also some small movement. Any periods of movement that are continuous for over 5 minutes, constitute *wakefulness*. The sum of all such periods of *wakefulness* during a sleep period, is called the *wake after sleep onset*, or *WASO*.

$$WASO = \sum_{\substack{Sleep\ Onset\ Time}}^{\substack{Sleep\ Awakening\ Time}} \begin{cases} ||WakefulPeriod||, & \text{if } ||WakefulPeriod|| > 5 \\ 0, & \text{otherwise} \end{cases} \tag{1.1}$$

Sleep Efficiency In sleep science, sleep quality is defined by a number of metrics, including total sleep time, wake after sleep onset, awakening index, and sleep efficiency [53]. There are also questionnaires that strive for a more subjective understanding of sleep quality. Since sleep efficiency combines many of the aforementioned metrics, it is generally accepted as the best objective measure of sleep quality. Thus, all work in this thesis treats sleep efficiency and sleep quality as synonymous. Sleep efficiency is computed as a numerical value ranging from 0 to 1. It is the ratio of total sleep time to total minutes in bed, i.e. the ratio of the length of the sleep period less the time spent awake, to the length of the sleep period plus the latency. According to specialists, a sleep efficiency below 0.85 (i.e., 85%) indicates poor sleep quality , and above 0.85 indicates good sleep quality [54].

$$\begin{aligned} Sleep\ Efficiency &= \frac{TotalSleepTime}{TotalMinutesinBed} \\ &= \frac{||SleepPeriod|| - WASO}{||SleepPeriod|| + Latency} \end{aligned} \tag{1.2}$$

1.7 Outline

This thesis creates novel, state-of-the-art, computer science algorithms for a translational science approach on improving human health through sleep quality. By providing the aforementioned tools for sleep and activity behaviour analysis, clinical decision-makers can deliver improved and informed healthcare. Moreover, this research also empowers patients from a quantified-self perspective, by conveying real-time recommendations to optimise productivity and improve quality of life. The goal is to empower clinicians and patients alike.

Chapter 2 focuses on screening. There is a massive market for better screening tools, as the majority of sleep conditions are left undiagnosed. This chapter surveys the ability of several statistical and data mining methods, to predict sleep efficiency proactively. It takes an in-depth look at analysing sensor time series with deep learning techniques for high fidelity prediction. It also introduces a novel deep leaning architecture called *time-batched long short-term memory cells* which builds off of the well-established long short-term memory cell spin-off from recurrent neural networks. This architecture shows a distinct improvement in its capability of handling the nature of wearable device data.

Chapter 3 moves on to the diagnosis stage of the sleep science process. Diagnosing a condition requires insights beyond what standard deep learning methods are capable of. This chapter introduces a feature construction method via a human activity recognition algorithm. This algorithm, RAHAR

(robust automated human activity recognition), provides a personalised activity labelling to a time series and a new level of automation that is not currently available in sleep research. Moreover, most human activity recognition algorithms aim to identify the *type* of behaviour. RAHAR is created to measure the *exertion level* of a behaviour.

Chapter 4 moves on to the therapy of sleep conditions. The objective of the work in this chapter is to provide recommendations for wearable device users to take action and improve their sleep quality. I define my methodology for extracting *behavioural recipes* through segmenting and clustering of the activity time series, in a retrospective manner. This novel way of building and evaluating a recommendation system allows for the rapid evolution of real-time sleep coaching.

Lastly chapter 5, highlights the main contributions of this body of work and takes a look into the future of the revised sleep science process. With the support of contemporary computational methods, there can be a profound improvement in society's quality of sleep and quality of life.

Chapter 2

Screening: A TB-LSTM Deep Learning Architecture

The National Institute of Health estimates that over 70 million Americans suffer from sleep problems, the majority of which are undiagnosed [55]. The current sleep science process has been limited to only the diagnosis and treatment of sleep conditions, but with the ubiquitous adoption of wearable devices, we are now entering a new era.

The *quantified self* movement has created a massive market for better screening tools. The individual no longer relies, or wants to rely, on a clinic visit to improve their health. However, widespread screening has not yet been feasible due to a major bottleneck in the analysis of actigraphy data. The evaluation and interpretation has required manual study by a clinical expert. Moreover, the process involves sleep experts manually configuring parameters prior to performing analysis. Automated computational methods can revolutionize this approach.

2.1 Approach

Introducing automation into the sleep science process is crucial for retrospective analysis. Section 2.2 details the system created in order to streamline the analysis of actigraphy data, by identifying the key time points illustrated in 2.7. This system could be integrated into any consumer wearable device or a clinically-validated device.

While retrospective analysis is useful in identifying whether or not the user endured a good or poor nights sleep, its conspicuous shortcoming is that it determines the quality of sleep after the fact. What is needed is a tool that can proactively predict the expected quality of sleep based on a user's activity. Recent systematic reviews have shown the relevance of physical activity to sleep, including sleep efficiency [56, 57, 58]. Although the relationship between physical activity and sleep is not yet fully understood, it is thought to be a strong and complex correlation.

Deep learning models have achieved state-of-the-art results in a wide variety of tasks in computer vision, natural language processing and speech recognition. The fact that deep learning models automatically learn abstract feature representations from raw features, while also optimizing on the target prediction tasks, makes them an attractive solution for predicting sleep quality from daily physical activity data.

The importance of this approach is two-fold:

- First, since our approach can be used in cases where sensory data during sleep is not available, our models can be used in the early detection of potential low sleep efficiency. This is a common problem with consumer-grade wearable devices, as users might not wear them during the night (battery recharging, sensors embedded in smart jewellery, and so forth).

- Second, our study was focused on advanced deep learning methods. Traditional prediction models applied to raw accelerometer data (e.g. logistic regression) suffer from at least 2 key

limitations: (*i*) They are not robust enough to learn useful patterns from noisy raw accelerometer output. As a result, existing methods for classification and analysis of physical activity rely on extracting higher-level features that can be fed into prediction models [59]. This process often requires domain expertise and can be time consuming. (*ii*) Traditional methods do not exploit task labels for feature construction, and thus can be limited in their ability to learn task-specific features. Deep learning has the advantage that it is robust to raw noisy data, and can learn, automatically, higher level abstract features by passing raw input signals through non-linear hidden layers while also optimizing on the target prediction tasks. This characteristic was leveraged by building models using a range of deep learning methods on raw accelerometer data. This reduced the need for data preprocessing and feature space construction and simplified the overall workflow for clinical practice and sleep researchers.

To my knowledge, this research is the first to look into the use of deep learning for the study of actigraphy data and the relationship between physical activity and sleep [60]. Previous research on the use of deep learning for sleep science has been focused on PSG data [61, 62].

2.2 Automated Actigraphy

Automating the actigraphy process is the first step to improving the scalability of sleep disorder screening. A fully automated actigraphy process would examine the individual's activity signal and determine the *Sleep Period*, *Sleep Onset*, *Awake Time*, *Time to Bed*, *Latency*, *WASO* and *Time to Active*.

The raw actigraphy output from a clinically-validated device such as the Agtigraph GT3X+, includes raw accelerometer output as well as a signal over time of *counts per minute*. These counts are the sum of band-pass filtered accelerometer output. The counts vary based on the frequency and intensity of the signal [63]. As mentioned, the counts are measured on a per minute basis, but

this can be altered to the desired epoch. Most consumer devices have the capability to collect at 30-100Hz. Figure 2.7 shows an example activity signal, i.e. *counts per minute* over time.

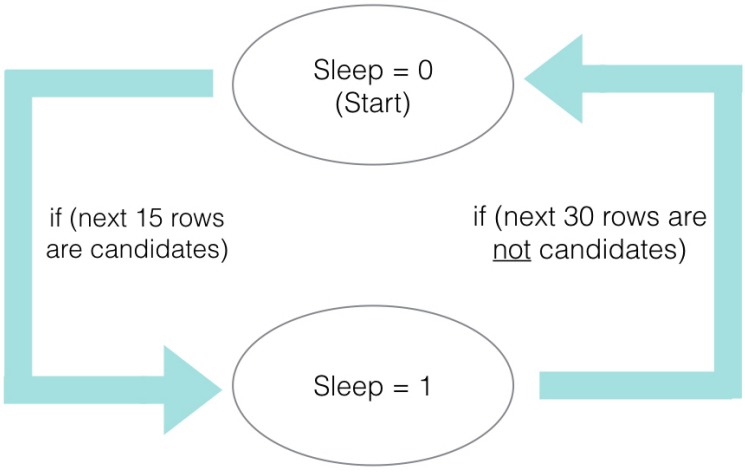
Once the activity signal is constructed, we must identify the sleep period. Traditional actigraphy generally includes a self-reported sleep diary that an individual fills in each night. The *time to bed* and the *time to active* is self-reported, and the *time to sleep* and *time to awake* is determined based on these boundaries. These records can be highly unreliable, especially in adolescents, so moving to an automated data-driven approach is again advantageous.

The first step is to run the entire raw accelerometer output through a state machine. Figure 2.1 illustrates the state machine design. The accelerometer output is triaxial, on an $\{x, y, z\}$ coordinate system. *Rows* refers to the aggregate accelerometer movement in an epoch of time, i.e. each epoch instance of $\{x, y, z\}$. *Candidate Rows* are rows with no triaxial movement, i.e. $\{x, y, z\}$=$\{0, 0, 0\}$, indicating the device is stationary. Note that depending on the device, there are different methods for detecting whether the device is being worn or not. The methods mentioned here are applicable to all. By definition of the state machine, an individual is asleep (*sleep onset time*) after 15 consecutive candidate rows, or 15 continuous minutes of being perfectly stationary. Note that a sedentary individual would not be able to hold perfectly still for this period of time. After 30 consecutive non-candidate row, or 30 minutes of at least minor continuous activity, the user is considered awake. This marks the *sleep awakening time*. The time period inbetween the *sleep onset time* and *sleep awakening time* is considered the *sleep period*.

WASO, or *wake after sleep onset*, is measured in the same way as defined in section 1.6. All moments of wakefulness are identified within the sleep period boundaries, as epochs with triaxial movement. If the wakefulness is continuous for over 5 minutes, it is aggregated into the WASO, otherwise it is discarded.

The *Time to Bed* and the *Latency* can be reverse engineered from the previous definitions with the use of human activity recognition. Chapter 3 elaborates on this process.

Figure 2.1: Automated Actigraphy State Machine



2.3 Deep Learning Methodology

In this section, the mathematics of the deep learning models (multi-layer perceptron, convolutional neural network, recurrent neural network, long short-term memory cells) used is described, as well as the specifics on the novel architecture I designed specifically for wearable data (time-batched long short-term memory cell).

2.3.1 Prediction Problem

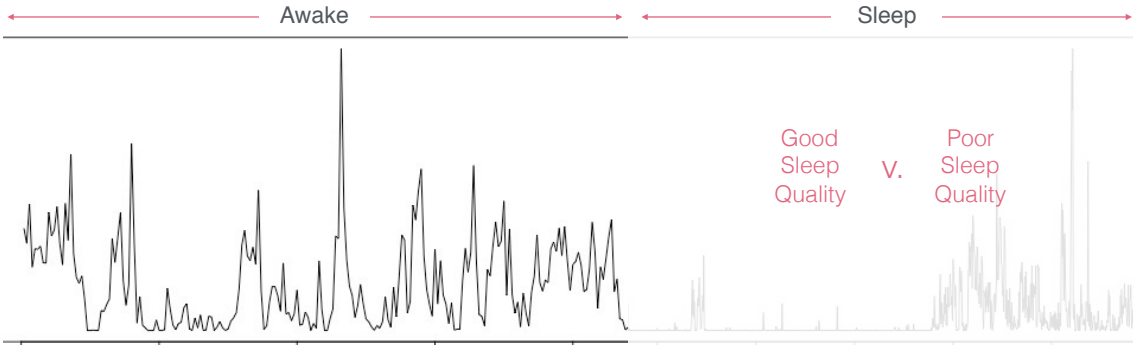
The input value to the prediction models is the activity signal, i.e. the counts per minute measured over time. The output is a binary classification of sleep quality as defined below. Figure 2.2 illustrates the use of an individual's activity signal to predict sleep quality.

The general concept of sleep quality naturally falls into the binary classifications of "good sleep" and "poor sleep". These are subjective measures, which is why many sleep science clinics use a

patient questionnaire to determine the sleep quality experienced. In the move to an automated, data-driven approach, I use sleep efficiency as the measure of sleep quality. Note that while sleep efficiency is a strong measure of sleep quality, latency is a strong measure of *perceived* sleep quality.

In order to evaluate the effectiveness of the models, the sleep efficiency as measured from the automated actigraphy system is used. The sleep community considers a sleep efficiency below 0.85 as poor quality sleep, and sleep efficiency of above 0.85. Thus this threshold is used to divide the binary classifications.

Figure 2.2: The Prediction Problem: Using Activity to Predict Sleep Quality



2.3.2 Deep Learning Modelling

Let $\mathbf{x_t} \in \mathbb{R}^D$ be a vector representing a person's activity measured at time t. Given a series of such input vectors $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ representing a person's physical activity in an awake time period, the deep neural models first compute compressed representations with multiple levels of abstraction by passing the inputs through one or more non-linear hidden layers. The abstract representations of the raw activity measures are then used in the output layer of the neural network to predict the sleep quality. Formally, the output layer defines a Bernoulli distribution over the sleep quality $y \in \{$*good, poor*$\}$:

$$p(y|\mathbf{X}, \theta) = \text{Ber}(y|\,\text{sig}(\mathbf{w}^{\mathbf{T}}\phi(\mathbf{X}) + b)) \tag{2.1}$$

where sig refers to the sigmoid function, $\phi(\mathbf{X})$ defines the transformations of the input $\mathbf{X}$ through non-linear hidden layers, $\mathbf{w}$ are the output layer weights and b is a bias term.

We train the models by minimizing the cross-entropy between the predicted distributions $\hat{y}_{n\theta} = p(y_n|\mathbf{X}_n, \theta)$ and the target distributions y_n (i.e., the gold labels).[1]

$$J(\theta) = -\sum_n y_n \log \hat{y}_{n\theta} + (1 - y_n) \log (1 - \hat{y}_{n\theta}) \tag{2.2}$$

Minimizing cross-entropy is same as minimizing the negative log-likelihood (NLL) of the data (or maximizing log-likelihood). Unlike generalized linear models (e.g., logistic regression), the NLL of a deep neural model is a non-convex function of its parameters. Nevertheless, we can find a locally optimal maximum likelihood (ML) or maximum a posterior (MAP) estimate using gradient-based optimization methods. The main difference between the models, as we describe below, is how they compute the abstract representation $\phi(\mathbf{X})$.

2.3.3 Multi Layer Perceptrons

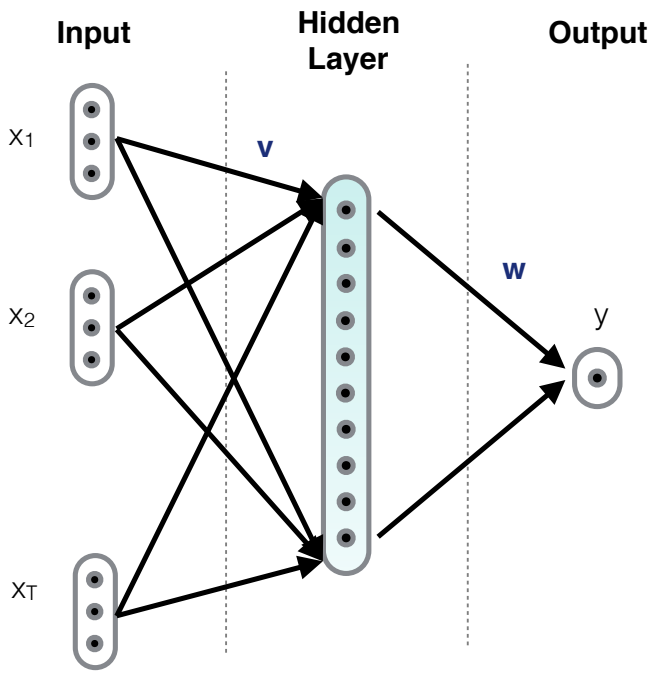
Multi-Layer Perceptrons (MLP), also known as Feed-forward Neural Networks, are the simplest models in the deep learning family. As shown in Fig. 2.3, the transformation of input $\phi(\mathbf{X})$ in MLP is defined by one or more fully-connected hidden layers of the form

$$\phi(\mathbf{X}) = f(V\mathbf{x}_{\mathbf{1}:\mathbf{T}}) = [f(\mathbf{v}_1^T\mathbf{x}_{\mathbf{1}:\mathbf{T}}), \cdots, f(\mathbf{v}_N^T\mathbf{x}_{\mathbf{1}:\mathbf{T}})] \tag{2.3}$$

where $\mathbf{x}_{\mathbf{1}:\mathbf{T}}$ is the concatenation of the input vectors $\mathbf{x}_1, \cdots, \mathbf{x}_T$, V is the weight matrix from the inputs to the hidden units, f is a non-linear activation function (e.g., $\text{sig}, \tanh$) applied element-wise,

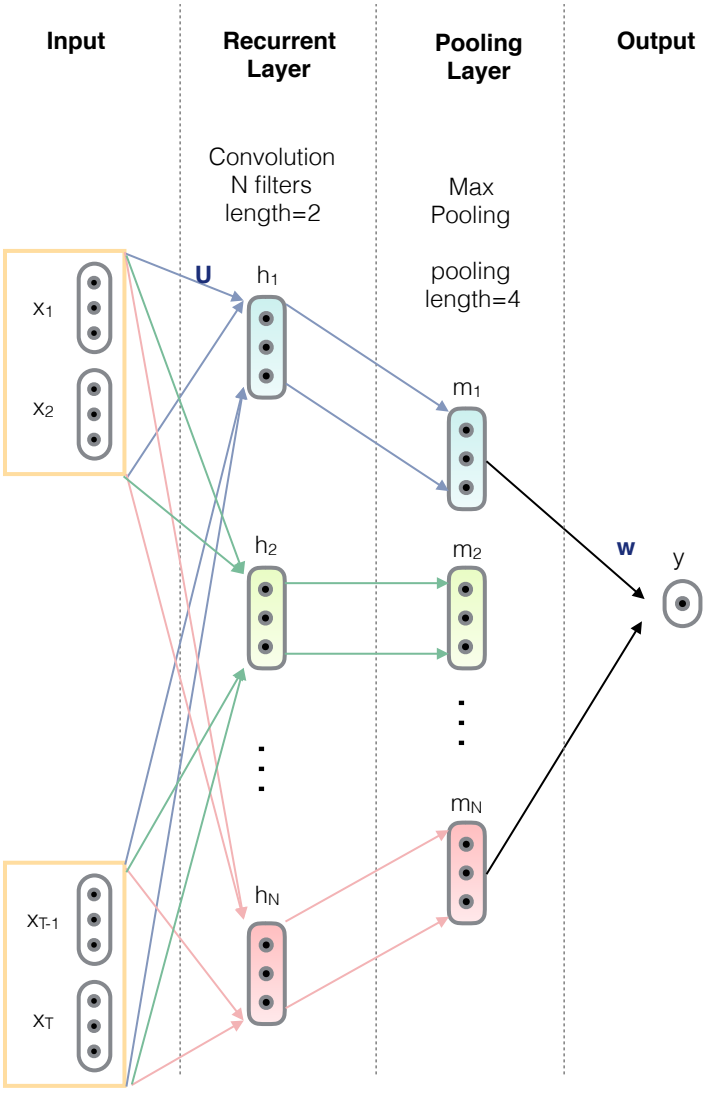
[1] Other loss functions (e.g., hinge) yielded similar results.

Figure 2.3: A Multi-Layer Perceptron with one hidden layer.



and N is the number of hidden units. MLP without any hidden layer (i.e., non-linear activations) boils down to a logistic regression (LR) or maximum entropy (MaxEnt) model. The hidden layers give MLP representational power to model complex dependencies between the input and the output.[2] By transforming a large and diverse set of raw activity measures into a more compressed abstract representation through its hidden layer, MLP can improve the prediction accuracy over linear models like LR.

Figure 2.4: A Convolutional Neural Network. Each coloured box in the second hidden layer represents a feature map, which is obtained by sliding its corresponding filter over the entire input. Pooling is independently done over feature maps.



2.3.4 Convolutional Neural Networks

The fully-connected MLP described above has two main properties: (*i*) it composes each higher level feature from the entire input sequence, and (*ii*) it is *time variant*, meaning it uses separate (non-shared) weight parameter for each input dimension to predict the overall sleep quality. However, a person's sleep quality may be determined by his activity over certain (local) time periods in the awake time as opposed to the entire awake time, and this can be invariant to specific timings. For example, high intensity exercises or games over certain period of time can lead to good sleep, no matter when the activities are exactly performed in the awake time. Furthermore, each person has his own habit of activities, e.g., some run in the morning while others run in the afternoon. A fully-connected structure would require a lot of data to effectively learn these specific activity patterns, which is rarely the case in health domain. Convolutional Neural Networks (CNN) address these issues of a fully-connected MLP by having repetitive filters or kernels that are applied to local time slots to compose higher level abstract features. The weights for these filters are shared across time slots.

As shown in Fig. 2.4, the hidden layers in a CNN are formed by a sequence of *convolution* and *pooling* operations. A convolution operation involves applying a *filter* $\mathbf{u}_i \in \mathbb{R}^{L.D}$ to a window of L input vectors to produce a new feature h_t

$$h_t = f(\mathbf{u}_i.\mathbf{x}_{t:t+L-1}) \tag{2.4}$$

where $\mathbf{x}_{t:t+L-1}$ denotes the concatenation of L input vectors and f is a non-linear activation function as defined before. We apply this filter to each window of L time steps in the sequence $\mathbf{X}$ to generate a *feature map* $\mathbf{h}_i = [h_1, \cdots, h_{T+L-1}]$. We repeat this process N times with N different filters to get N different feature maps $[\mathbf{h}_1, \cdots, \mathbf{h}_N]$. Note that we use a *wide* convolution rather than a *narrow*

[2]In fact, MLP has been shown to be a *universal approximator*, meaning that it can model any suitably smooth function to any desired level of accuracy, given the required hidden units [64].

one, which ensures that the filters reach the entire sequence, including the boundary slots [65]. This is done by performing *zero-padding*, where out-of-range ($t<1$ or $t>T$) slots are assumed to be zero.

After the convolution, we apply a max-pooling operation to each feature map

$$\mathbf{M} = [\mu_p(\mathbf{h}_1), \cdots, \mu_p(\mathbf{h}_N)] \tag{2.5}$$

where $\mu_p(\mathbf{h}_i) = \mathbf{m}_i$ refers to the max operation applied to each window of p features in the feature map $\mathbf{h}_i$. For $p = 2$, this pooling gives the same number of features as in the feature map (because of the zero padding).

Intuitively, the filters compose activity measures in local time slots into higher-level representations in the feature maps, and max-pooling reduces the output dimensionality while keeping the most important aspects from each feature map. Since each convolution-pooling operation is performed independently, the features extracted become invariant in locations, i.e., when they occur in the awake time. This design of CNNs yields fewer parameters than its fully-connected counterpart, therefore, generalizes well for target prediction tasks.

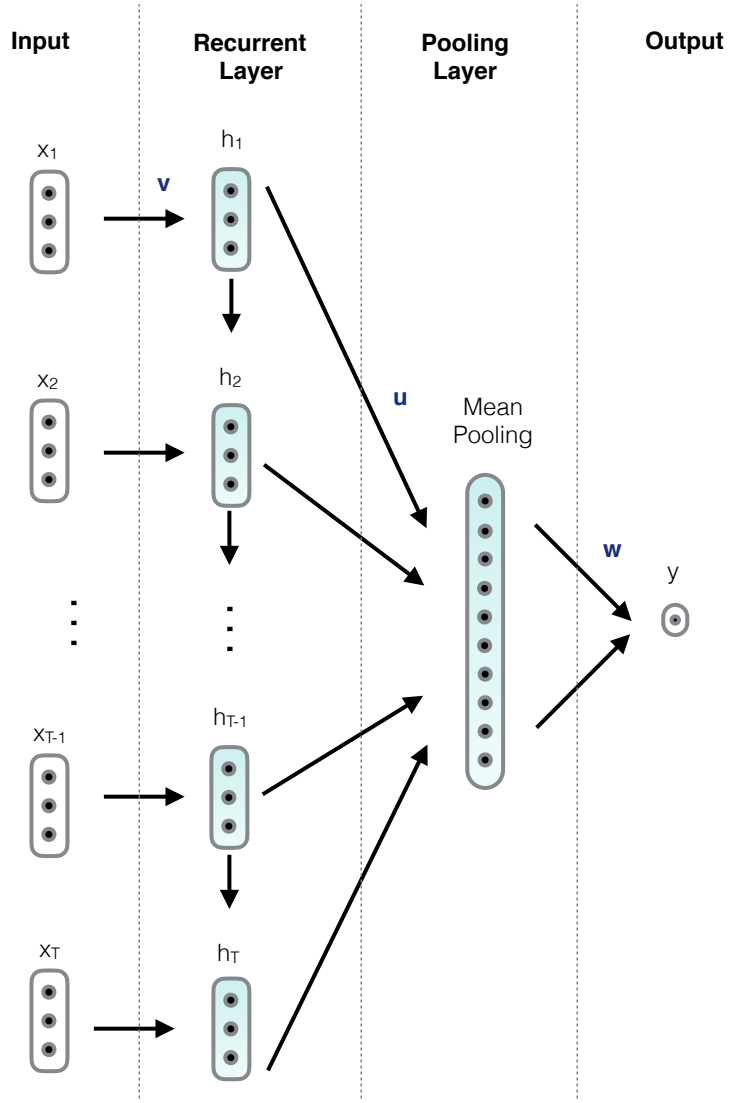
2.3.5 Recurrent Neural Networks

In CNN, features (or attributes) are considered in a bag-of-slots fashion disregarding the order information. The order in which the activities were performed in an awake time could be important. Recurrent Neural Networks (RNN) compose abstract features by processing activity measures in an awake time sequentially, at each time step combining the current input with the previous hidden state. More specifically, as depicted in Fig. 2.5, RNN computes the output of the hidden layer $\mathbf{h}_t$ at time t from a non-linear transformation of the current input $\mathbf{x}_t$ and the output of the previous hidden layer $\mathbf{h}_{t-1}$. More formally,

$$\mathbf{h}_t = f(U\mathbf{h}_{t-1} + V\mathbf{x}_t) \tag{2.6}$$

where f is a non-linear activation function as before, and U and V are compositional weight matrices. RNNs create internal states by remembering previous hidden layer, which allows them to exhibit dynamic temporal behaviour. We can interpret $\mathbf{h}_t$ as an intermediate representation summarizing the past. The representation for the entire sequence can be obtained by performing a pooling operation (e.g., *mean-pooling, max-pooling*) over the sequence of hidden layers or simply by picking the last hidden layer $\mathbf{h}_T$. In our experiments, we found mean-pooling to be more effective than other methods.

Figure 2.5: A Recurrent Neural Network with one recurrent layer.



RNNs are generally trained with the backpropagation through time (BPTT) algorithm, where errors (i.e., gradients) are propagated back through the edges over time. One common issue with

BPTT is that as the errors get propagated, they may soon become very small or very large that can lead to undesired values in weight matrices, causing the training to fail. This is known as the *vanishing* and the *exploding* gradients problem [66]. One simple way to overcome this issue is to use a truncated BPTT [67] for restricting the backpropagation to only few steps like 4 or 5. However, this solution limits the simple RNN to capture long-range dependencies. Below we describe an elegant RNN architecture to address this problem.

2.3.6 Long Short-Term Memory Cell Recurrent Neural Networks

Long Short-Term Memory or LSTM [68] is specifically designed to capture long range dependencies in RNNs. The recurrent layer in a standard LSTM is constituted with special units called *memory blocks* (Figs. 2.5 and 2.6). A memory block is composed of four elements: (*i*) a memory cell c (a neuron) with a self-connection, (*ii*) an input gate i to control the flow of input signal into the neuron, (*iii*) an output gate o to control the effect of the neuron activation on other neurons, and (*iv*) a forget gate f to allow the neuron to adaptively reset its current state through the self-connection. The following sequence of equations describe how the memory blocks are updated at every time step t:

$$\mathbf{i}_t = \text{sigh}(U_i\mathbf{h}_{t-1} + V_i\mathbf{x}_t + \mathbf{b}_i) \tag{2.7}$$

$$\mathbf{f}_t = \text{sigh}(U_f\mathbf{h}_{t-1} + V_f\mathbf{x}_t + \mathbf{b}_f) \tag{2.8}$$

$$\mathbf{c}_t = \mathbf{i}_t \odot \tanh(U_c\mathbf{h}_{t-1} + V_c\mathbf{x}_t) + \mathbf{f}_t \odot \mathbf{c}_{t-1} \tag{2.9}$$

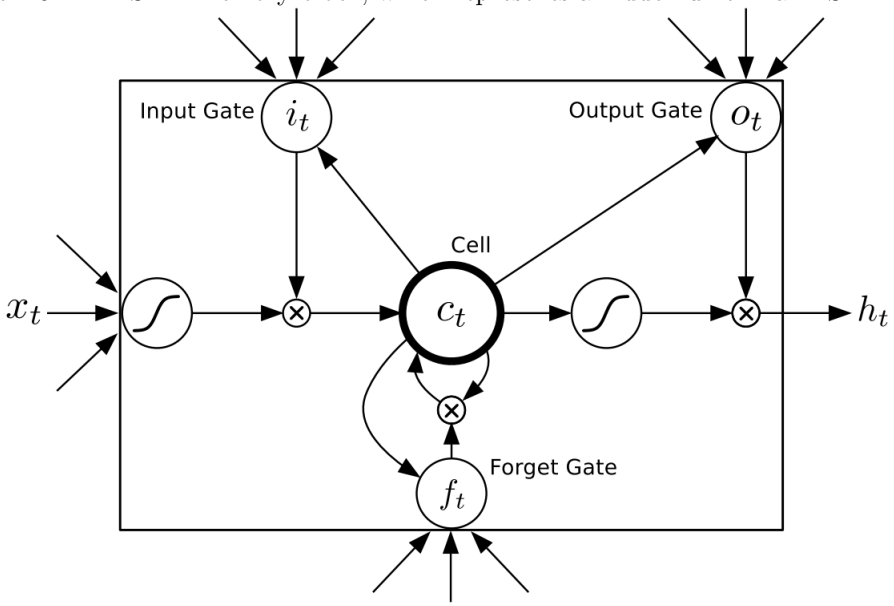
$$\mathbf{o}_t = \text{sigh}(U_o\mathbf{h}_{t-1} + V_o\mathbf{x}_t + \mathbf{b}_o) \tag{2.10}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{2.11}$$

where U_k and V_k are the weight matrices between two consecutive hidden layers, and between the input and the hidden layers, respectively, which are associated with gate k (input, output, forget and

cell); and and $\mathbf{b}_k$ is the corresponding bias vector. The symbols sigh and tanh denote hard sigmoid and hard tan, respectively, and the symbol $\odot$ denotes a element-wise product of two vectors. LSTM by means of its specifically designed gates (as opposed to simple RNNs) is capable of capturing long range dependencies.

Figure 2.6: An LSTM memory block, which represents a hidden unit in an LSTM-RNN.



2.3.7 Time-batched Long Short-Term Memory Cell Recurrent Neural Networks

The naive way to apply RNNs to activity records is to consider the accelerometer output aggregated at every minute in an awake time as a separate time step. In our data, each time step contains one measure (i.e. the vertical axis). This results in very lengthy sequences with only one feature at each time step. RNNs applied to this setting suffer from two problems: (i) because of the low dimensional input at each time step, RNNs become ineffective in composing features and fail to

capture the sequential dependencies, and (*ii*) because of long sequences, the gradients (errors) from the final time step vanish before they reach to earlier parts of the network, causing the training with BPTT to fail even with LSTM cells.

To circumvent this problem, we construct batches of time steps by merging accelerometer measures over S time steps. In other words, each time step in the batch setting generates an input vector of $S \times D$ elements. To make all input vectors equal-sized, we use zero-padding for the last time step (if needed). Note that setting $S = 1$ gives the original sequences.

2.4 Experimental Design

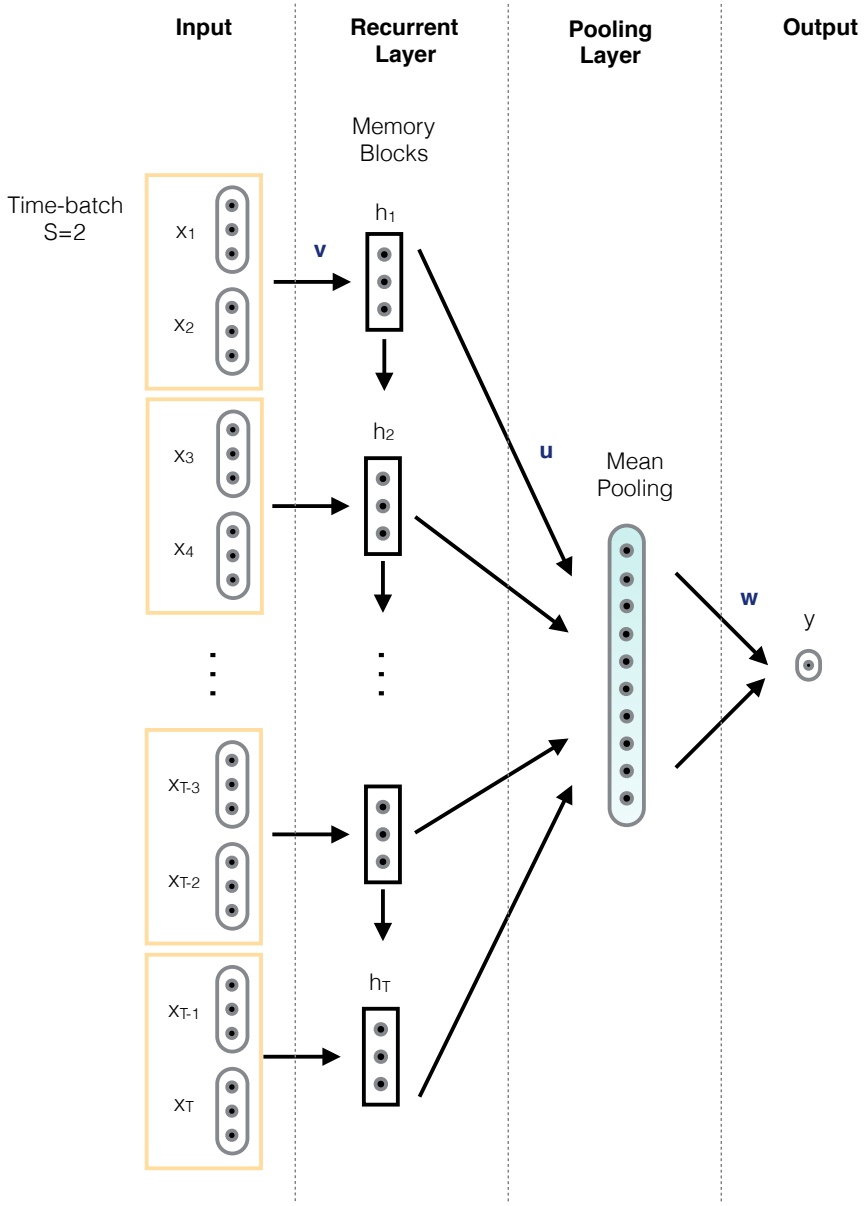
In this section, the experimental design, details on the data used, and experimental results are described.

2.4.1 Data Collection

The data used in this experiment was collected as part of a research study to examine the impact of sleep on health and performance in adolescents by Weil Cornell Medical College - Qatar. Two international high schools were selected for cohort development. The participants consisted of 92 male and female adolescents from ages 10 to 17.

Student volunteers were provided with an actigraph accelerometer, ActiGraph GT3X+1, to wear on their non-dominant wrist continuously throughout the one-week observational trial. The device was water resistant and fully charged so the device would not need to be removed for any reason (i.e. even when sleeping and showering). De-identified data collected in the study were used in this analysis. The wearable device sampled the users sleep-wake activity at 30-100 Hertz. Currently sleep experts use this device in conjunction with the accompanying software, ActiLife [63], to evaluate an individuals sleep period. We evaluate our results side-by-side with ActiLifes results.

Figure 2.7: Time-Batched Long Short-Term Memory Cell Recurrent Neural Network Architecture



2.4.2 Data Partitioning

To train the models without over-fitting and test their performance afterwards, a random partitioning of the dataset was created. Each time series was assigned to a partition randomly while maintaining an even class distribution of the target variable, sleep quality. The data were split with a 70%-15%-15% ratio for training, testing, and validation sets, respectively. All reported results are based on model predictions on the test set.

2.4.3 Data Staging

The input of the models were time series vectors, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, representing the physical activity of a persons awake time. Each vector corresponded to a continuous period of awake time, and so for each individual, there might be multiple such vectors over the 7 days. Each $\mathbf{x}_T$ represented the triaxial value of the vertical axis at time $\mathbf{t}$.

The output of the model was a binary classification decision between good and poor sleep quality based on the sleep efficiency equation (%). These classifications were determined by using automated actigraphy to evaluate the sleep quality retrospectively, and were used as the ground truth. Good quality sleep corresponds to a sleep efficiency above 0.85, and poor quality sleep corresponds to an efficiency below or equal to 0.85. In addition to the binary decision, the model also gave its confidence (a score between 0.0 and 1.0) in that decision.

2.4.4 Model Training

To be able to predict, the models were first trained on the training dataset, using an online training algorithm, RMSprop [69], which relied on a number of preset parameters:

- Dropout ratio: The ratio of hidden units to turn off in each mini-batch training.
- Mini-batch size: The number of training instances to consider at one time.

- Learning rate: The rate at which the parameters are updated.
- Max epoch: The maximum number of iterations over the training set.

The training algorithm minimizes the cross-entropy between the predicted distribution and the actual (ground truth) target labels. To avoid over-fitting, we used early stopping based on the models performance on the validation set. In particular, the model was evaluated after every epoch on the validation set and stopped when its accuracy went down. To reduce the cross-entropy between the predicted distributions and the target distributions, RMSprop was used setting the maximum number of epochs to 50

Rectified linear units (ReLU) were used for the activation functions (f). Dropout [70] of hidden units was also used to avoid overfitting. Regularization on weights such as l_1 and l_2 did not improve the results. We experimented with $DR \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ dropout rates and $MB \in \{5, 10, 15, 20\}$ minibatch sizes. For MLP, we experimented with one hidden layer containing $N \in \{2, 3, 5, 10, 15, 20\}$ units. For CNN, we experimented with $N \in \{25, 50, 75, 100, 125, 150\}$ number of filters with filter lengths $FL \in \{2, 3, 4, 5\}$ and pooling lengths $PL \in \{2, 3, 4, 5\}$. For simple and LSTM RNNs, we experimented with $N \in \{25, 50, 75, 85, 100\}$ hidden units in the recurrent layer, and $S \in \{25, 50, 75, 100\}$ time-slots for constructing batches in the time-batched LSTM. Since the size of the training data is small, the network weights are initialized with zero to start the training with the simplest model. Optimal model configurations for each setting is summarized in table 2.1 and the following subsections.

Logistic Regression

As a baseline model , we used logistic regression (LR) to predict the sleep quality. LR is a generalized linear classification model that does not have any hidden layers. For the LR, the raw input signals, X, are directly fed into the output layer for prediction without any non-linear hidden layer transformations. The optimal setting for logistic regression (LR) was with a mini-batch size of 5

Table 2.1: Optimal Model Configurations

Model	MLP*	LR	MLP	CNN	RNN	LSTM	TB-LSTM
Dropout Ratio	0.3	0.5	0.1	0.0	0.1	0.5	0.5
Mini-Batch Size	5	5	20	5	5	5	5
Hidden Layer Size	15	-	15	-	75	100	100
Time Batch Size	-	-	-	-	-	-	50
Number of Filters	-	-	-	25	-	-	-
Filter Length	-	-	-	5	-	-	-
Pool Length	-	-	-	4	-	-	-

* indicates MLP used with RAHAR output

and a dropout ratio of 0.5.

Multi Layer Perceptron

MLPs, also known as feed-forward neural networks, are the simplest models in the deep learning family. They have one or more hidden layers. In fact, MLP without any hidden layers is equivalent to logistic regression. In MLP, all the units of a hidden layer are fully connected to the units in the previous layer. The best parameter configuration for MLP was with a mini-batch size of 20, a dropout ratio of 0.1, and a hidden layer size of 15.

Convolutional Neural Network

CNNs are a more complex type of deep learning method that includes repetitive filters or kernels applied to local time slots, thereby composing a high level of abstract features. This operation is called convolution. After convolution, a max-pooling operation is performed to select the most significant abstract features. This design of CNNs yields fewer parameters than its fully connected counterpart (MLP), and therefore generalizes well for target prediction tasks. For its best configuration, we used 25 hidden nodes, filter length of 5 and pooling length of 4, 5 mini-batch size, and 0.0 dropout ratio.

Recurrent Neural Network

RNNs compose abstract features by processing activity measures in an awake time sequentially, at each time step combining the current input with the previous hidden state. RNNs create internal states by remembering the previous hidden layer, which allows them to exhibit dynamic temporal behavior. These features make RNNs a good deep learning method for temporal series. RNNs performed best with a mini-batch size of 5, a dropout ratio of 0.1 and a hidden layer size of 75.

Long Short-Term Memory Cell Recurrent Neural Network

A subtype of RNN, LSTM uses specifically designed memory blocks as units in the recurrent layer to capture longer-range dependencies. The optimal configuration values for LSTM were a mini batch size of 5, dropout ratio of 0.5, and hidden layer size of 100.

Time-Batched Long Short-Term Memory Cell Recurrent Neural Network

To further improve our implementation of LSTM, we constructed batches of time steps by merging accelerometer measures over time steps. We referred to this version of the model as TB-LSTM. The optimal configuration values for TB-LSTM were mini-batch size of 5, dropout ratio of 0.5, and hidden layer size of 100.

2.4.5 Evaluation

For the evaluation of the performance of the different models, several well-known metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve (AUC), are used. These metrics are commonly used in data mining and clinical decision support systems.

Accuracy

It is computed as the proportion of correct predictions, both positive and negative (sum of true positives and true negatives divided by the number of all instances in the dataset).

Precision

It is the fraction of the number of true positive predictions to the number of all positive predictions (i.e. true positives divided by the sum of true positives and false positives). In our case, precision described what percentage of the time the model predicted good-quality sleep correctly. Note that precision is also known as positive predictive value.

Specificity

It is the fraction of the number of true negative predictions to the actual number of negative instances in the dataset (i.e. true negatives divided by the sum of true negatives and false positives). In our case, specificity referred to the percentage of the correctly predicted poor-quality sleep to the total number of poor-quality sleep instances in the dataset. Note that specificity is also known as true negative rate.

Recall or Sensitivity

It is the fraction of the number of true positive predictions to the actual number of positive instances in the dataset (i.e. true positives divided by the sum of true positives and false negatives). In our case, recall referred to the percentage of the correctly predicted good-quality sleep to the total number of good-quality sleep instances in the dataset. Note that recall is also known as true positive rate or sensitivity.

F1-Score

There is usually an inverse relationship between precision and recall. That is, it is possible to increase the precision at the cost of decreasing the recall, or vice versa. Therefore, it is more useful to combine them into a single measure such as F1 score, which computes the harmonic mean of precision and recall.

Area Under the ROC Curve

It represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Hence, AUC defines an effective and combined measure of sensitivity and specificity (which are often inversely related, just like precision and recall) for assessing inherent validity of a classifier.

2.5 Results

As shown in table 2.2 and figure 2.8, the performance of the logistic regression in the metrics previously explained, performed worse than the models based on deep learning. Only the simple RNN performed worse than logistic regression in both F1-score (harmonic mean of precision and recall) and accuracy.

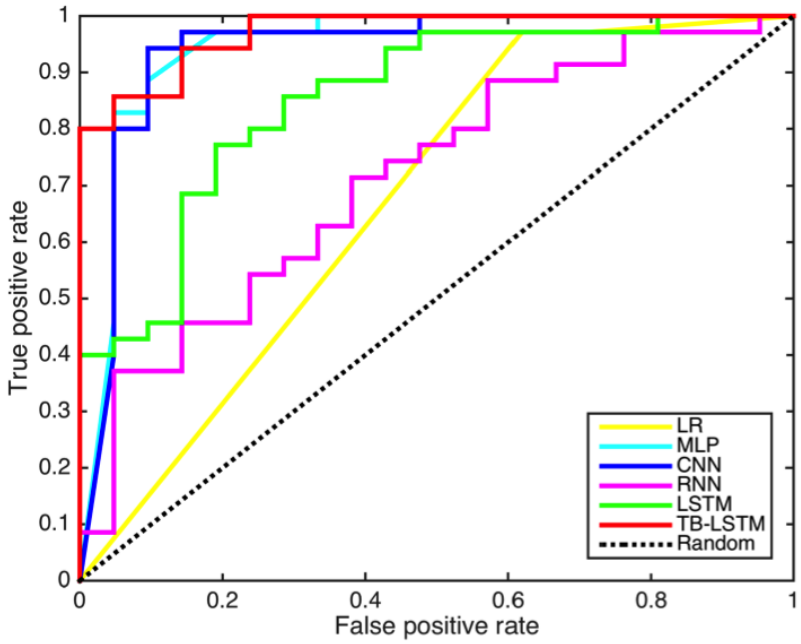
Also shown in table 2.2, the AUC of the logistic regression model was low. The AUC value for logistic-regression was 0.6463, which was close to 0.5 (equivalent to a random prediction). This showed the limitation of classical models in analysing raw accelerometry.

In contrast, all the AUC values for the deep learning models were better with a range from 0.7143 to 0.9714, TB-LSTM being the best performer and RNN, the worst. Time-batched LSTM, CNN, and MLP performed the best with AUC scores showing an improvement over LR by 50%, 46%, and 46%, respectively.

Table 2.2: Deep Learning Results on Raw Accelerometer Data

	AUC	F1-Score	Precision	Recall	Accuracy
LR	0.6463	0.8193	0.7083	0.9714	0.7321
MLP	0.9449	0.9118	0.9394	0.8857	0.8929
CNN	0.9456	0.9444	0.9189	0.9714	0.9286
RNN	0.7143	0.7711	0.6667	0.9143	0.6607
LSTM	0.8531	0.8500	0.7556	0.9714	0.7857
TB-LSTM	0.9714	0.9211	0.8537	1.0000	0.8929

Figure 2.8: ROC curves of logistic regression (LR), multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory cell (LSTM), and the novel time-batched long short-term memory cell (TB-LSTM).



2.5.1 Comparison Between Deep Learning Models

Upon comparing the deep learning neural network models, it is apparent that CNN yielded slight improvement over MLP in AUC (0.07% absolute), but more in F1 (3.57%) and accuracy (4.00%).

These improvements could be attributed to the time- invariant convolution-pooling operations of the CNN model to pick key local patterns, which generalized well for small training data. The F1-score improved by 12% for time-batched LSTM, by 15% for CNN, and by 11% for MLP. The accuracy improved by 22% for time-batched LSTM, by 27% for CNN, and by 22% for MLP.

A comparison of the RNN models revealed that LSTM outperformed simple RNN by a wide margin; 19%, 10%, and 19% in AUC, F1, and accuracy, respectively. These gains over simple RNNs could be attributed to the specially designed gates of LSTMs that could capture long-range dependencies between physical activities in the sequences.

However, this is not surprising. Both simple and LSTM RNNs operate on sequences, where each time step comprises only one activity value. This often results in very long sequences. As mentioned earlier, in this setting RNNs cannot compose higher-level features effectively because of low-dimensional input at each time step, and also suffer from vanishing gradient problems due to lengthy sequences.

The solution to surmount this problem was to use a time-batched input to LSTM. When comparing the results of the novel time-batched LSTM (TB-LSTM) with those of MLP and CNN, the TB-LSTM outperformed both MLP and CNN in AUC by 3%; in fact, it had the highest AUC score. It achieved better F1 score than MLP (1%), but worse than CNN (2%). When observing their precision and recall values, TB-LSTM had a very high recall but lower precision, which meant that it tended to predict more *good quality sleep* than the gold standard. For the same reason, its accuracy was also lower than that of CNN.

2.6 Discussion

2.6.1 Principal Findings

This study focused on the prediction of poor versus good sleep efficiency. It is a simple, but important, problem as sleep efficiency has been found to be a crucial sleep parameter with important health consequences [38,43,44]. Furthermore, the prediction did the estimate the overall sleep efficiency but rather the differentiation between two classes (poor versus good sleep efficiency). This classification is consequently not an indicator of sleep patterns, but the prediction of a sleep quality parameter that might indicate a potential sleep problem.

As in prediction or diagnostic problems, the results need to be discussed in terms of sensitivity and specificity (see table 2.3). The deep learning methods of CNN and TB-LSTM were the best performers overall. Their sensitivity (0.97 and 1, respectively) showed that these models were able to detect nearly all the cases of good-quality sleep, meaning that in a tool for screening potential sleep problems these models will be able to detect easily people with normal sleep quality. Often high sensitivity comes at the price of low specificity (i.e. failing to identify negative cases, or true negative error). This was the case of logistic regression, which had a high sensitivity but a specificity of 0.3, meaning that in such models many poor sleeps would have been wrongly classified as good sleep. This is very important, since misidentifying poor sleep cases can lead to under-diagnosis of problematic sleep.

The sensitivity (also known as recall) and specificity of each of the models are reported in table 2.3. The high sensitivity values of each of the models indicate that deep learning has a strong capability of correctly identifying individuals with good sleep patterns from their preceding awake activity. The specificity is high for TB-LSTM, MLP, and CNN, indicating that these models were also able to successfully distinguish those with poor sleep patterns.

Table 2.3: The Sensitivity and Specificity of the Deep Learning Models

	Sensitivity	Specificity
LR	0.9714	0.3333
MLP	0.8857	0.9048
CNN	0.9714	0.8571
RNN	0.9143	0.2381
LSTM	0.9714	0.4762
TB-LSTM	1.0000	0.7143

2.6.2 Impact on Sleep Science and eHealth

Millions of consumers are purchasing wearable devices that incorporate activity sensors. This burst of human activity data is a great opportunity for health research, but to achieve this paradigm shift, it is necessary to develop new algorithms and tools to analyse this type of data. Furthermore, sleep insufficiency is highly prevalent in contemporary society, and has been shown to influence energy balance by altering metabolic hormone regulation. Consequently, health researchers are exploring the impact of sleep and physical activity on many health conditions. A major bottleneck for this research is that current approaches for studying actigraphy data require intensive manual work by human experts. Moreover, huge datasets of actigraphy data are emerging from health research, including the study of sleep disorder patients, healthy populations, and epidemiological studies.

As explained in the results, this research supports the feasibility of using physical activity data to predict the quality of sleep in terms of sleep efficiency. Improved algorithms, such as the ones presented in this study, can lead to a paradigm shift in the study of lifestyle behaviours such as sleep and physical activity, just as electrocardiography became crucial for cardiology and clinical research.

This work provides an early example of how advanced deep learning methods can be used to

infer new insights from raw actigraphy data. The results illustrate that deep learning performed better than classical methods in terms of learning useful patterns from raw accelerometer data for the task of sleep quality prediction.Since deep learning models compute abstract features from raw input signals while optimizing on the actual sleep quality, this process yields a more robust solution. Furthermore, the good results of deep learning show that raw accelerometer data had more signal regarding sleep quality, which traditional models such as logistic regression are not able to capture.

The focus on predicting and forecasting can help design new eHealth applications where predictions are made to personalize coaching for patients or to facilitate decision making of professionals.More research needs to be done to understand why deep learning performs better, which eventually can help in identifying new factors influencing the quality of sleep.

There is an increased interest in sleep in the health domain. This is consequently being reflected in an increased use of eHealth for sleep [71, 72, 73, 31, 74] and also in the use of social media for sharing sleep logs [31]. The expansion of eHealth into sleep is not limited to sleep disorders, but also to improve sleep for people living with chronic conditions such as cancer. These developments are closely related to the concept of Quantified Self for health [75, 76]. Furthermore, it can be assumed that predicting sleep quality based on physical activity data acquired by accelerometer data (both from actigraphy or activity trackers) can be used to provide personalized feedback, such as momentary ecological interventions based on mobile technology [77].

The deep learning algorithms predicted a parameter regarding the quality of sleep solely relying on the physical activity during the awake time. To my knowledge, this has not been previously done. The advantage of this approach is that eventually the same technique can be used to predict sleep quality with data from smart watches and other wearable devices that are not necessarily used during sleep. Therefore, our models can eventually be used within eHealth applications that do not require wearing a sensor during sleep. This is of special interest for the development of smart watch health applications [78], as they might require frequent battery charging

The application of this research has many limitations as explained in the next subsection. However, this is the first study, to focus on the prediction of sleep quality from physical activity accelerometer data during awake periods. The methodology and results can be used as the baseline for further studies looking into predicting sleep quality from mobile and wearable devices. This is a source of major concern, since many sleep applications, predict sleep quality with unclear methodology and performance [71, 73, 72]. Although more studies are highlighting the increasing reliability of consumer sleep wearables [79], we do not know how they calculate or predict sleep quality parameters. To maximize the potential of ubiquitous wearables for sleep health, we need research on not only the reliability of consumer-grade devices but also their data processing and modelling techniques.

2.6.3 Limitations

There are some limitations in this work regarding the generalization of the sleep outcomes. Sleep behaviour can be affected by cultural aspects and may also change with age. This study sample drew sleep data from a cohort of adolescents, aged 10-17 years, living in the capital of Qatar, Doha. Future research will need to evaluate whether applications of deep learning for sleep research using actigraphy will yield similar results in different populations, for example, adults or people with chronic conditions.

In terms of the novel architecture designed, the prediction problem was simplified to good and poor sleep quality with regards to sleep efficiency. This may be an oversimplification of complex sleep problems or wearable device analysis in other domains. To provide more precise predictions (e.g. quantitative value of sleep efficiency), these techniques will need to be expanded for regression. Moreover, a prediction, such as the one presented in this study, might be useful for the detection of people with unhealthy sleep patterns, but not to identify the causes of poor sleep efficiency.

There is also a limitation in the interpretation of deep learning. Deep learning models are black

boxes and do not provide explanation of their prediction. Other techniques such as logistic regression can provide insights on which features contribute to the prediction. However, this study showed that the performance of such models was much lower than that of deep learning. Nevertheless, new techniques in deep learning are being researched to facilitate the interpretation of such models.

Lastly, in this study, the models used the data from the individuals awake time to predict sleep quality. The prediction was made at the last moment before sleep, using the full awake time activity. If these models were to be used to provide personalized feedback to individuals with sleep problems, they will need to be tested with fragments of the awake time, giving an individual time to alter their behaviour. Since our data was fragmented into sleep periods and awake times specific to an individual, the models are generalizable to handle varying durations of awake time. This problem is addressed in more detail in chapter 4.

2.7 Conclusion

These chapter has provided 2 key techniques to radically update the sleep science process. The first technique was the automation of actigraphy analysis. This system eliminates the sleep clinic bottle neck and has the potential to greatly reduce the number of undiagnosed sleep conditions, by alerting users of serious problems. It also allows the consumer with an interest in optimizing their sleep, to evaluate their activity with their own wearable device.

The second contribution was the ability to make a proactive sleep quality prediction based on an individual's daily activity. The results show the feasibility of deep learning in predicting sleep efficiency using wearable data from awake periods. Our novel architecture showed an ability to interpret the specific low-dimensional and longitudinal nature of wearable device data. This is of particular importance because deep learning eliminates the need for data preprocessing and simplifies the overall workflow in clinical care and sleep research. The feasibility of this approach can lead to new applications in sleep science and also to the development of more complex eHealth sleep

applications for both professionals and patients. These models can also be integrated in the broader context of quantified self.

Chapter 3

Diagnosis

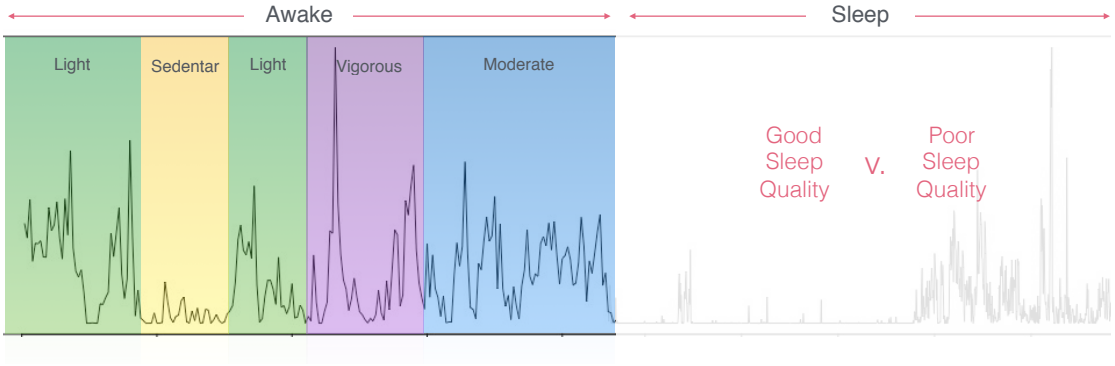
The biggest shortcoming of the deep learning tools introduced in the last chapter, is the lack of transparency. Deep learning tools are 'black boxes' in that do not explain why a decision or prediction was made. Although there are many researchers working on 'unboxing' the algorithmically determined rules, this research is still preliminary. "Black box" results are particularly problematic for some of the disciplines that artificial intelligence and deep learning are expected to revolutionize. These tasks (such as high-stake trading decisions, loan approval, autonomous driving, etc) can lead to very serious consequences. In fact, in 2018 the European Union may require companies to justify decisions made my autonomous systems [80].

In the case of sleep screening, determining whether or not a sleep problem is present, is the first step. The second step requires diagnosing the condition. While the deep learning methods provided a high-fidelity prediction model for sleep quality, they provided no insights or information into why that sleep quality was occurring. The models automatically learnt abstract feature representations from the raw accelerometer input. In this chapter, a method for creating interpretable feature representations is presented. The goal of these representations is threefold:

- To improve the prediction ability of downstream analysis using traditional data mining methods, such as support vector machines and decision trees. These methods struggle to perform on the raw accelerometer data.
- To provide insights and interpretability into the rationale behind the predictions. Traditional data mining modelling will allow for this.
- To advance automated actigraphy to the next level, and provide personalized measurements of an individual's activity.

Currently, the sleep science process uses polysomnography, home sleep tests and actigraphy to diagnose sleep conditions. All of these methods require in clinic visits, and rely on clinical experts to manually evaluate the data. The automated actigraphy process described in section 2.2, identifies the key actigraphy milestones and replaces this cumbersome process. However, this process can be taken one step further by the incorporation of human activity recognition algorithms. Figure 3.1 illustrates the usage of human activity recognition features for sleep quality prediction.

Figure 3.1: The Prediction Problem: Using human activity recognition as a feature representation



3.1 Approach

Human Activity Recognition (HAR) is the understanding of human behaviour from data captured by pervasive sensors, such as cameras or wearable devices. It is a powerful tool in medical application areas, where consistent and continuous patient monitoring can be insightful. Wearable devices provide an unobtrusive platform for such monitoring, and due to their increasing market penetration, feel intrinsic to the user. This daily integration into a users life is crucial for increasing the understanding of overall human health and well-being.

Wearables, such as actigraph accelerometers, generate a continuous time series of a persons daily physical exertion and rest. This ubiquitous monitoring presents substantial amounts of data, which can (*i*) provide new insights by enriching the feature set in health studies, and (*ii*) enhance the personalisation and effectiveness of health, wellness, and fitness applications. By decomposing an accelerometers time series into distinctive activity modes or actions, a comprehensive understanding of an individuals daily physical activity can be inferred. The advantages of longitudinal data are however complemented by the potential of noise in data collection from an uncontrolled environment. Therefore, the data sensitivity calls for robust automated evaluation procedures.

In this chapter, a robust automated human activity recognition (RAHAR) algorithm [81] is presented. The algorithm is tested in the application area of sleep science by providing a novel framework for evaluating sleep quality and examining the correlation between the aforementioned and an individuals physical activity. Even though the performance of the proposed HAR algorithm is evaluated on sleep analysis, RAHAR can be employed in other research areas such as obesity, diabetes, and cardiac diseases.

3.2 Human Activity Recognition

Human activity recognition (HAR) has been an active research area in computer vision and machine learning for many years. A variety of approaches have been investigated to accomplish HAR ranging from analysis of still images and videos to motion capture and inertial sensor data.

Video has been the most widely studied data source in HAR literature. Hence, there exists a wealth of papers in this particular domain. The most recent literature on HAR from videos include trajectory-based descriptors [82, 83, 84], spatio-temporal feature representations [85, 86, 87], feature encoding [88, 89, 90], and deep learning [91, 92, 93]. Reviewing the extensive list of video-based HAR studies, however, goes beyond the scope of this study and we refer the reader to [94, 95] for a collection of more comprehensive surveys on the topic.

Unlike HAR from video, existing approaches for HAR from still images are somewhat limited, and range from histogram based representations [96, 97] and colour descriptors [98] to pose-based, appearance-based and parts-based representations [99, 100, 101, 102]. Guo and Lai recently provided a comprehensive survey of the studies on still image-based HAR in [103].

Several techniques have been proposed, on the other hand, for HAR from 3D data, encompassing representations based on bag-of-words [104, 105], eigen-joints [106], sequence of most informative joints [107], linear dynamical systems [108], actionlets [109], Lie algebra embedding [110], covariance descriptors [111], hidden Markov models [112], subspace viewinvariant metrics [113] and occupancy patterns [114, 115]. Aggarwal and Xia presented a recent survey summarizing state-of-the-art techniques in HAR from 3D data [116].

Unlike vision-based HAR systems, sensor-based HAR technologies commonly deal with time series of state changes and/or various parameter values collected from a wide range of sensors such as contact sensors, accelerometers, audio and motion detectors, etc. Chen et al. [117] and Bulling et al. [118] present comprehensive reviews of sensor-based activity recognition literature. The most recent work in this domain includes knowledge-based inference [119, 120], ensemble methods

[121, 122], data-driven approaches [123, 124], and ontology-based techniques [125].

All of the aforementioned studies investigate recognition/classification of fully observed action or activity, e.g., jumping, walking, running, drinking, etc. (i.e., activities of daily living), using well-curated datasets. However, thanks to the quantified self movement, myriad of consumer-grade wearable devices have become available for individuals who have started monitoring their physical activity on a continuous basis, generating tremendous amount of data. Therefore, there is an urgent need for automatic analysis of data coming from fitness trackers to assess the physical activity levels and patterns of individuals for the ultimate goal of quantifying their overall well-being. This task requires understanding of longitudinal, noisy physical activity data at a rather higher (coarser) level than specific action/activity recognition level. Main challenges as well as opportunities of HAR from personalized data and life-logs have been discussed in several dimensions in [126, 127, 128, 129, 130].

There has been a number of initiatives to overcome the challenge of collecting annotated personalized data to further research on HAR from continuous measurement of real-world physical activities [131, 132]. Even though such systems exhibit a crucial attempt in furthering research in mining personalized data, they have limited practical importance as they rely on manual annotation of the acquired data. There has also been recent attempts to automatically recognize human activities from continuous personalized data [133, 134, 135, 136]. However, most of these studies are designed to recognize only a predefined set of activities, and hence, not comprehensive and robust enough to quantify the physical activity levels for the overall assessment of individuals well-being.

3.3 Methodology

The methodology for RAHAR is shown algorithmically in figure 3.2. We elaborate on the details of the algorithm in this section.

Figure 3.2: An Illustration of the RAHAR Workflow

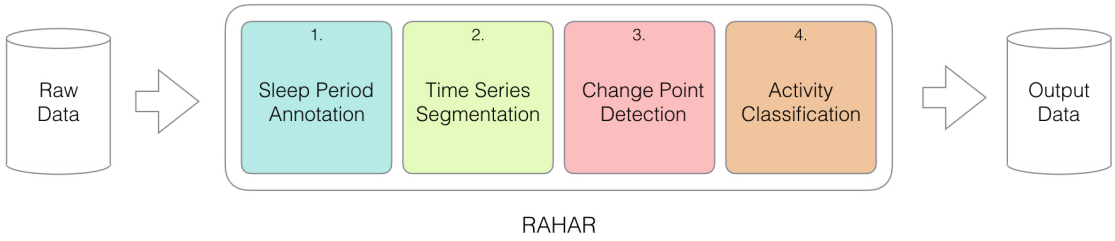
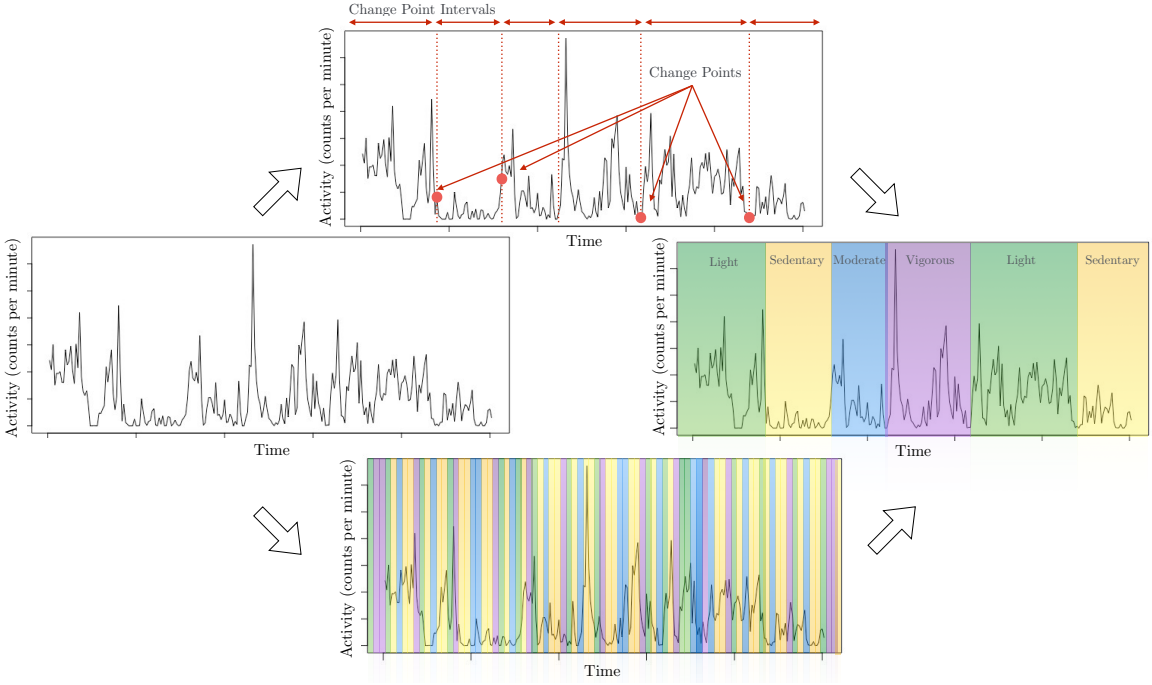


Figure 3.3 provides a high level overview of key steps in the RAHAR pipeline. The right most graphic illustrates the robust and automated human activity recognition labels on the time series.

Figure 3.3: High-level overview of RAHAR



3.3.1 Sleep Period Annotation

The sleep period annotation phase builds on top of automated actigraphy. However, since RAHAR can be used as a stand alone tool for human activity recognition in any domain, it is not assumed that automated actigraphy is completed, and we start with the raw, non-annotated activity signal. The accelerometer output is put through the state machine in figure 2.1. The details of this process are in section 2.2.

The key outputs from the state machine are the *sleep onset time* and *sleep awakening time*. These terms border the *sleep period*, but in reverse, also border the awake time of an individual. As shown in figure 3.1, RAHAR identifies exertion levels of a individual through their awake time, so identifying this time period is critical.

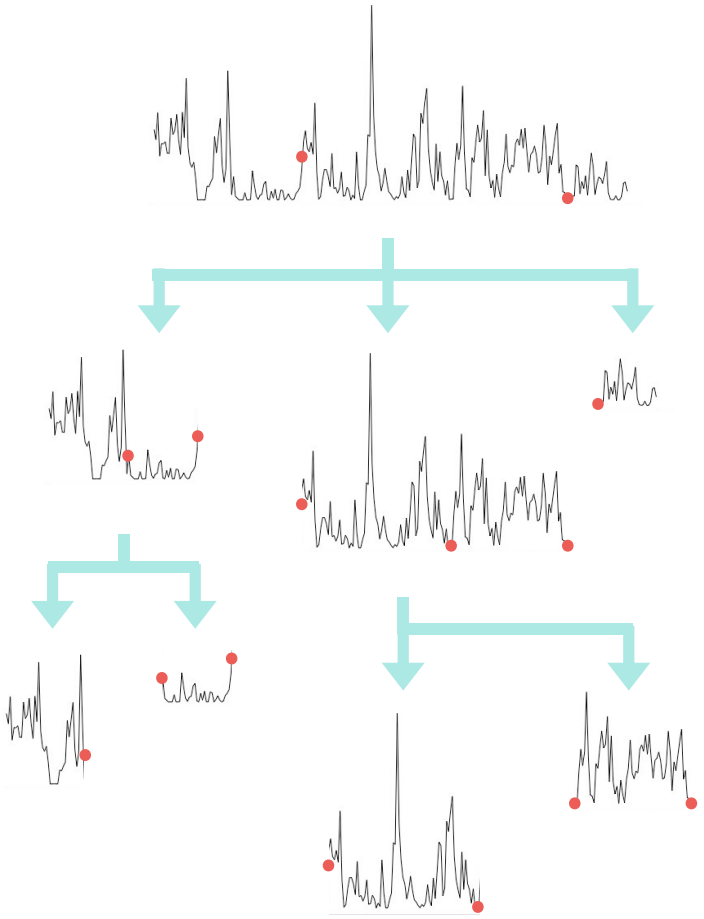
3.3.2 Time Series Segmentation

Since wearable device data is collected continuously, the time series can carry on over multiple awake and sleep periods. For this reason, the time series is divided into multiple awake-time-sleep-time pairs, by segmenting the time series at every awake time. Since RAHAR can be used as a diagnostics tool for sleep quality prediction, the human activity recognition determinations in the preceding awake time, are used to predict the following sleep period. This process is described in further detail in the evaluation section.

3.3.3 Change Point Detection

One of the key components of RAHAR is its personalisation. RAHAR does not create arbitrary divisions in activity (e.g. cut-offs every hour), or create population level generalizations, (e.g count per minutes thresholds). Rather it uses, change point detection to find localized distributional changes within the local time series that indicate a change in behaviour. This is done for each individual time series using a method called hierarchical divisive estimation.

Figure 3.4: A Visualisation of Change Point Detection Using Hierarchical Divisive Estimation



Hierarchical divisive estimation can be broken down into three fundamental steps: (*i*) measure the difference in the distribution, (*ii*) estimate the location of the change points hierarchically, and (*ii*) test the significance of the change point. The mathematical details of how an individual change point is found are described below. For further details please refer to the original paper, [137]. Figure 3.4 illustrates how the change points are identified hierarchically. Upon each division in

the time series, points are located where the distribution differs. These points are then tested for statistical significance. If the point is statistically significant, it is considered a change point, and the time series is divided. This process continues iteratively in a hierarchical manner, until no further statistically significant change points can be identified.

Let $x_1, x_2, \cdots, x_T \in R$ be the activity signal of counts per minute over time, as seen in figure 3.5. Define $Y_n = \{Y_i : i = 1, 2, \cdots, n\}$ and $Z_m = Z_j : j = 1, 2, \cdots, m$. This gives the sample divergence measure as:

$$\hat{\varepsilon}(Y_n, Z_m(\kappa); \alpha) = \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} |Y_i - Z_j|^\alpha - \binom{n}{2}^{-1} \sum_{1 \leqslant i < k \leqslant n} |Y_i - Y_k|^\alpha - \binom{m}{2}^{-1} \sum_{1 \leqslant j < k \leqslant m} |Z_j - Z_k|^\alpha \tag{3.1}$$

where the scaled empirical divergence is:

$$\hat{\mathcal{Q}}(Y_\tau, Z_\tau(\kappa); \alpha) = \frac{mn}{m+n} \hat{\varepsilon}(Y_\tau, Z_\tau(\kappa); \alpha) \tag{3.2}$$

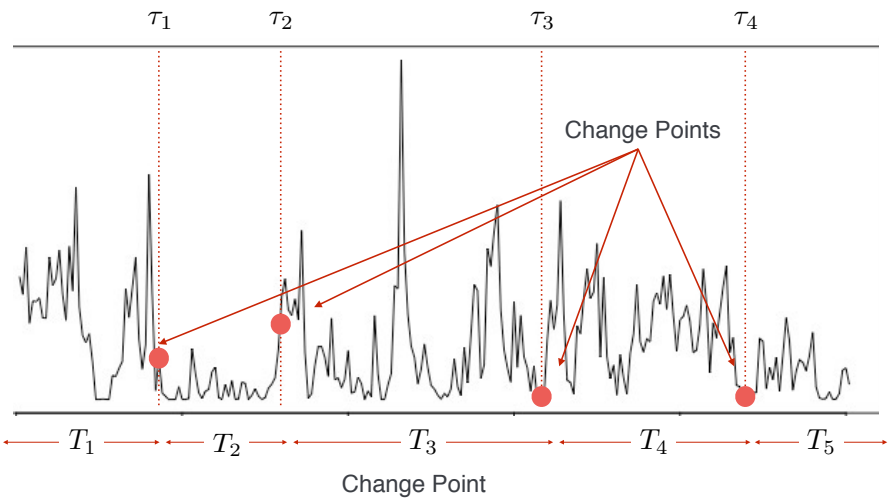
Defining, $Y_\tau = x_1, x_2, \cdots, x_\tau$ and $Z_\tau(\kappa) = x_{\tau+1}, x_{\tau+2}, \cdots, x_\kappa$ where $1 \leqslant \tau < \kappa \leqslant T$, the change point is then:

$$(\hat{\tau}, \hat{\kappa}) = \underset{(\tau, \kappa)}{\operatorname{argmax}} \hat{\mathcal{Q}}(Y_\tau, Z_\tau(\kappa); \alpha) \tag{3.3}$$

As mentioned, the segment is tested to find a point that could divide the time series such that the division would create 2 vectors with maximum divergence. For the purposes of implementation, a maximum number of random permutations of 99 was selected.

The purpose of identifying the change points is to identify the key moments in an activity time series, where there is a quantitative change in behaviour. Figure 3.5 shows how the activity time series would be segmented. Each change point makes the boundaries of a change-point-interval.

Figure 3.5: An Illustration of *change points* and *change point intervals*



3.3.4 Activity Classification

The activity signal from the accelerometer data contains post-filtered counts for each of the $\{x, y, z\}$ axes. These counts quantify the frequency and intensity of the users activity. Using Troianos cut point scale [138], the age of a user, and their accelerometer triaxial count, each epoch in the time series can be labelled with an intensity level: sedentary, light, moderate, or vigorous activity.

In this case, each epoch is one minute in length. Interpreting an individual's behaviour at a minute-by-minute level provides an unnecessary granularity. Moreover, it is highly subject to noise. Aggregating the cut point activity intensity labels, over the change point intervals, smoothens the signal and provides a more robust learning representation. To determine the exertion level of each change point interval, RAHAR calculates the statistical mode of the cut point labels over each epoch.

Not only does this process personalize the exertion level interpretation, but it automates the activity labelling, eliminating the need for cumbersome annotation by sleep experts, and enabling the widespread evaluation required in large clinical trials.

3.4 Experimental Design

RAHAR is fundamentally a feature extraction algorithm for human activity recognition, in the context of quantifying daily physical activity exertion levels of individuals. Although RAHAR can be used in other domains, it is tested here as a tool for sleep science diagnosis. As a result, the evaluation of RAHAR should include an assessment of its output, used as input into a prediction model for sleep quality. As opposed to the algorithms used in chapter 2, using RAHAR, will provide insights into the behaviour of an individual.

The best evaluation, is to test the quality of activity recognition by RAHAR as compared to the current process, i.e. an expert using actigraphy software to label the exertion levels. There is no ground truth on human activity in this context, in other words there is no way of knowing the reality of what an individual exerted. Instead, the objective is to evaluate which approach leads to better higher-fidelity models for predicting sleep quality.

Four models were selected for evaluating the performance of RAHAR against the performance of an expert-annotated time series: logistic regression, support vector machines with radial basis function kernel, random forest, and adaboost.

- Logistic Regression (LogR): This model was chosen because it is an easily interpretable binary classifier. It is also relatively robust to noise, which as explained earlier is a complication on data collected in an uncontrolled environment. [Note: Even though logistic regression (LogR) is included in the experiments, it is important to note that LogR model failed to stratify the dataset successfully for the state-of-the-art baseline approach, and predicted all cases to be in a single class. Therefore, the LogR score of RAHAR is excluded from analysis whenever corresponding LogR score of the state-of-the-art baseline approach were not available.]

- Support Vector Machine (SVM): This model was selected because it, also, is a binary classifier. using a radial basis function kernel, it differs from logistic regression in that it does not linearly

divide the data.

- Random Forest (RF): This model was tested as an alternative because of its easy straightforward interpretation, which is particularly relevant in the healthcare or consumer domains. It also is not restricted to linearly dividing the data.
- Adaboost (Ada): Lastly, Adaboost was tested because it is less prone to overfitting than random forest.

3.5 Results

Table 3.1 and 3.2 disclose the results of using RAHAR on the data described in section 4.3.1, with the different prediction models. 'SE+AL' refers to a sleep expert using the ActiLife software, as opposed to the automated process of RAHAR.

Table 3.1: RAHAR Prediction Evaluation Metrics

	AU-ROC		F1 Score		Accuracy		Precision	
	SE+AL	RAHAR	SE+AL	RAHAR	SE+AL	RAHAR	SE+AL	RAHAR
Ada	0.7489	0.8132	0.5574	0.6885	0.6966	0.7206	0.5667	0.9130
RF	0.8115	0.8746	0.6885	0.7500	0.7865	0.7647	0.7000	0.9231
SVM	0.7497	0.7895	0.3721	0.7077	0.6966	0.7206	0.6667	0.8519
LogR	0.5884	0.8649	-	0.6875	-	0.7059	-	0.8462
Average	0.7246	0.8355	0.5393*	0.7154*	0.7266*	0.7353*	0.6445*	0.8960*

* logistic regression (LogR) score is not included in averaging.

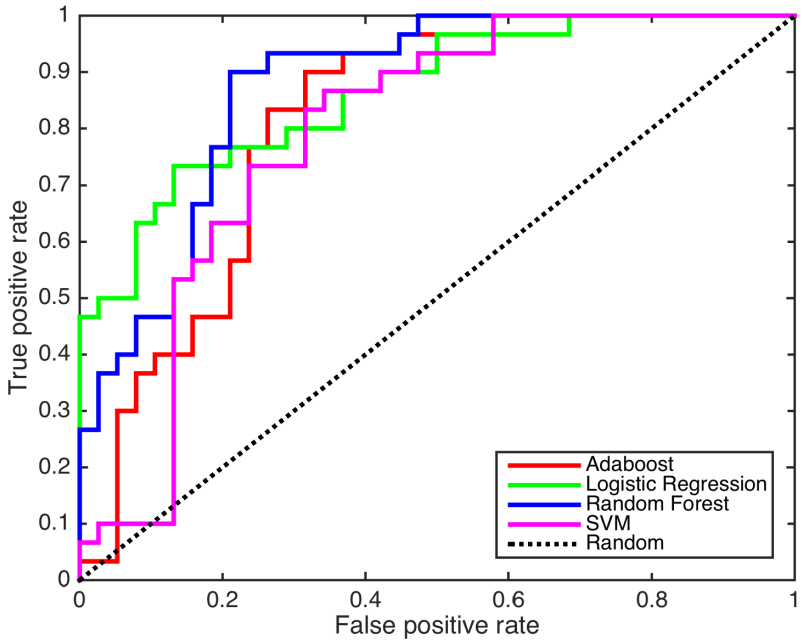
Table 3.2: RAHAR Health Evaluation Metrics

	Recall / Sensitivity		Specificity	
	SE+AL	RAHAR	SE+AL	RAHAR
Ada	0.5484	0.5526	0.7759	0.9333
RF	0.6774	0.6316	0.8448	0.9333
SVM	0.2581	0.6053	0.9310	0.8667
LogR	-	0.5789	-	0.8667
Average	0.4946*	0.5965*	0.8505*	0.9111*

* logistic regression (LogR) score is not included in averaging.

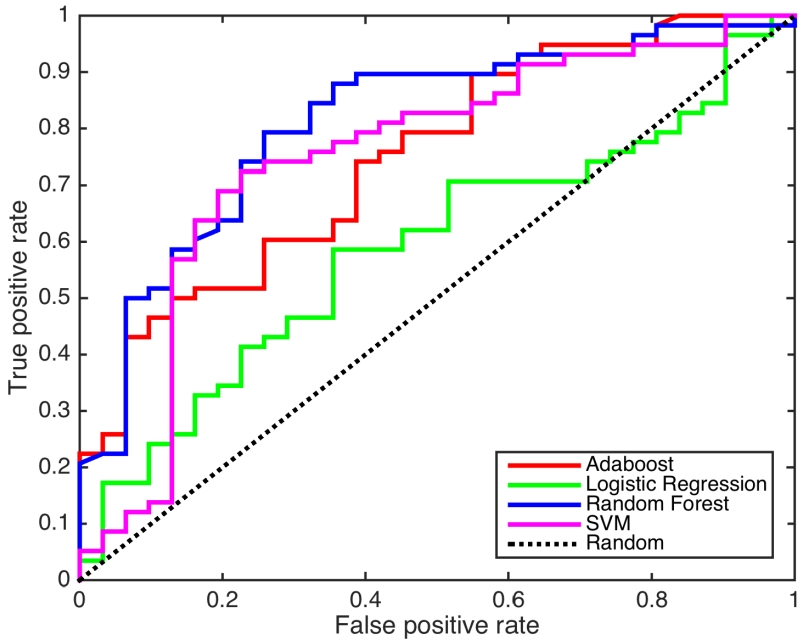
Figures 3.6 and 3.6 show the ROC curves for RAHAR and the sleep expert using ActiLife software (denoted as SE+AL), respectively. Table 3.1 summarises the results for both RAHAR and SE+AL. One of the most important performance measures for human activity recognition is the area under the ROC curve (AU-ROC). Based on AU-ROC scores, both RAHAR and SE+AL performed best using random forest as the prediction model. Furthermore, SE+AL achieved an average AU-ROC of 0.7246 whereas RAHAR achieved 0.8355. This shows a 15% improvement of AU-ROC score, on average, by using RAHAR instead of placing the burden of manual analysis on a sleep expert. With an AU-ROC score of 0.5884 for SE+AL approach, the logistic regression model was unable to stratify the dataset, and so predicted all cases to be in a single class. This is a failure of the logistic regression model for this problem, and thus, its results are not included in this discussion. For this reason, the misleading results have also been removed from table 3.1.

Figure 3.6: ROC Curves for RAHAR



Another important performance measure for human activity recognition algorithms is the F1 score. It is computed as the harmonic mean of precision and recall. According to table 3.1 and table 3.2, RAHAR performed better than SE+AL in terms of precision and recall for all models, and hence, yielded significantly higher F1 scores. Specifically, the F1 score for RAHAR, on average, was 0.7154 whereas it was 0.5393 for SE+AL (excluding logistic regression in both cases), yielding a solid margin of about 0.18 points (i.e., more than 30% improvement). On the other hand, the accuracy scores, on average, were 0.7353 for RAHAR and 0.7266 for SE+AL (again excluding logistic regression), and exhibited a relatively less significant difference still in favour of RAHAR.

Figure 3.7: ROC Curves for Sleep Expert using ActiLife software



3.6 Discussion

In this section we discuss the results of the best performing model and its broader impact to the area of sleep science. As seen in figure 3.6 and , random forest and logistic regression were the two best performing models with the RAHAR algorithm. Based on the desired threshold value of true positive rate, TPR, (i.e., sensitivity), either model could be preferred to minimize false positive rate, FPR, (i.e., 1-specificity), which is equivalent to maximizing specificity. Random forest was also the best performing model for the SE+AL approach as mentioned earlier. If we compare the ROC as well as the sensitivity-specificity plots of the best model of each approach (i.e., random forest), we see that RAHAR outperforms SE+AL almost always as illustrated in figure 3.8 and 3.9.

Figure 3.8: Comparison of the ROC for RAHAR and SE+AL on the Best Model

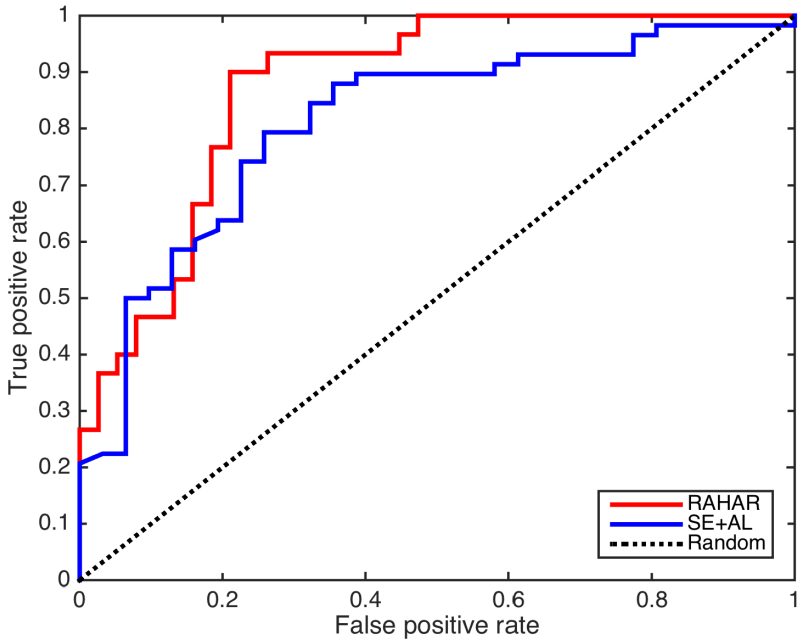
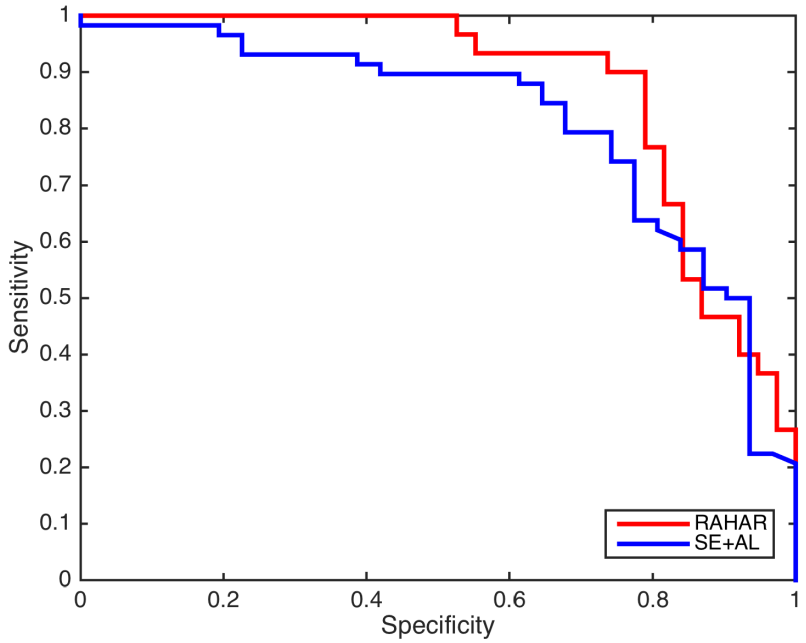


Table 3.2, on the other hand, summarises sensitivity and specificity scores for RAHAR and SE+AL. The average sensitivity score for SE+AL and RAHAR across all models except logistic regression were 0.4946 and 0.5965, respectively. In other words, average sensitivity score for RAHAR is 20higher than that of SE+AL. As for specificity, RAHAR with an average score of 0.9111 outperforms SE+AL with an average score of 0.8505, which corresponds to a 7% improvement.

As we seek to determine whether an individual had good quality sleep based on their physical activity levels during awake period prior to sleep, a false positive occurs when the model predicts good quality sleep when the person actually had a poor quality sleep. Therefore, the number of false positives needs to be kept at a minimum for a desired number of true positives. In other words, a high specificity score is sought after while keeping the sensitivity score at the desired level. As

can be seen from figure 3.9 with this perspective in mind, for a large range of sensitivity scores, RAHAR achieved higher specificity scores almost all the time than SE+AL did. For example, RAHAR achieved a sensitivity score of 0.9 with a specificity score of 0.8 whereas SE+AL remained at a specificity score of 0.6 for the same sensitivity threshold.

Figure 3.9: Comparison of Sensitivity and Specificity for RAHAR and SE+AL on the Best Model



In summary, RAHAR outperforms the current state-of-the-art procedure in sleep research. However, its application is not limited to sleep and it can be used for the understanding and treatment of other health issues such as obesity, diabetes, or cardiac diseases. Moreover, RAHAR allows for fully automated interpretation without the necessity of manual input or subjective self-reporting.

3.7 Conclusion

This chapter presented a robust automated human activity recognition (RAHAR) algorithm for multi-modal phenomena, and evaluated its performance in the application area of sleep science diagnostics. The results of RAHAR were tested against the results of a sleep expert using ActiLife. The goal was to predict sleep quality, specifically, sleep efficiency. RAHAR, (*i*) automated the activity recognition, and (*ii*) improved the current state-of-the-art results, on average, by 15% in terms of AU-ROC and 30% in terms of F1 scores across different models. Automating the human activity recognition puts sleep science evaluation in the hands of wearable device users. This empowers users to self-monitor their sleep-wake habits, and take action to improve the quality of their life. The improved results demonstrate the robustness of RAHAR as well as the capabilities of implementing the algorithm within clinical software such as ActiLife.

Chapter 4

Therapy: A Dynamic Activity Recommendation System

Minor sleep conditions are sleep patterns that result in fatigue, discomfort, or personal dissatisfaction. As aforementioned, therapy for these minor sleep conditions can include medication and dietary supplements or population level guidelines. More serious sleep conditions, such as sleep apnea and insomnia, require clinical care. The therapeutic approaches for those conditions can include surgery, oral appliances, or take home equipment.

Medication and dietary supplements are a form of therapy for people suffering from minor sleep conditions. However the sleep medication market amounts to over $2.1 billion, and only provides a temporary suppressing of the symptoms [50]. Another key component is lifestyle change. But these guidelines are population level suggestions that are impersonal, and in today's *internet of things* and *quantified self* environment, outdated. Society is ready for more.

While there are a plethora of sleep tracking applications and devices on the market, few provide recommendations. Whilst awareness of sleep quality is an important step in the right direction,

educating individual's on their poor sleep quality without providing recommendations to improve it, can be discouraging and detrimental. Thus, the impact of the research in this chapter is immediate and in high-demand.

The only other dynamic recommendation systems that are deployed today, are traffic applications such as Google Maps, Waze, Apple Maps etc. These applications frequently update the recommended driving route based on ongoing traffic, as well as the frequent change in a person's location. Developing an activity recommendation system however, is a unique challenge due to the distinct nature of physical behaviour.

In this chapter, the design, methodology, and retrospective evaluation for a dynamic and real-time activity recommendation system is described. This recommendation system could be used as a sleep coach within either a quantified self application, or as an automated assistant to cognitive behavioural therapy interventions.

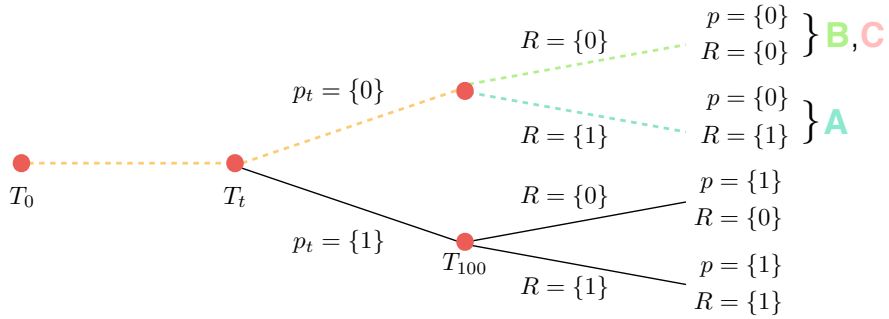
4.1 Approach

All the prediction modelling techniques used in the previous chapters, provide insights into sleep quality before it occurs, but *immediately* before it occurs, allowing for no behavioural adjustment. In order for the activity recommendation system described in this chapter to succeed, it needs to first predict the *expected* sleep quality at any given time t (i.e. p_t) rather than predicting at the end of the day (i.e. p_{100}, where 100 is the percentage of total awake time). This prediction will be critical in determining what recommendation to give. If good sleep is predicted, no intervention is necessary, or a user may be encouraged to continue similar behaviour. If poor sleep is predicted, a variety of possible recommendations may be provided to remedy the activity deficit.

Figure 4.1 illustrates a two-dimensional tree of the different possible predicted versus real outcomes. Activity recommendations for behaviour change would be given to individuals who are predicted to have poor sleep ($p_t = \{0\}$). This would result in a real sleep quality (R) that is either

good ($R = \{1\}$) or poor ($R = \{0\}$).

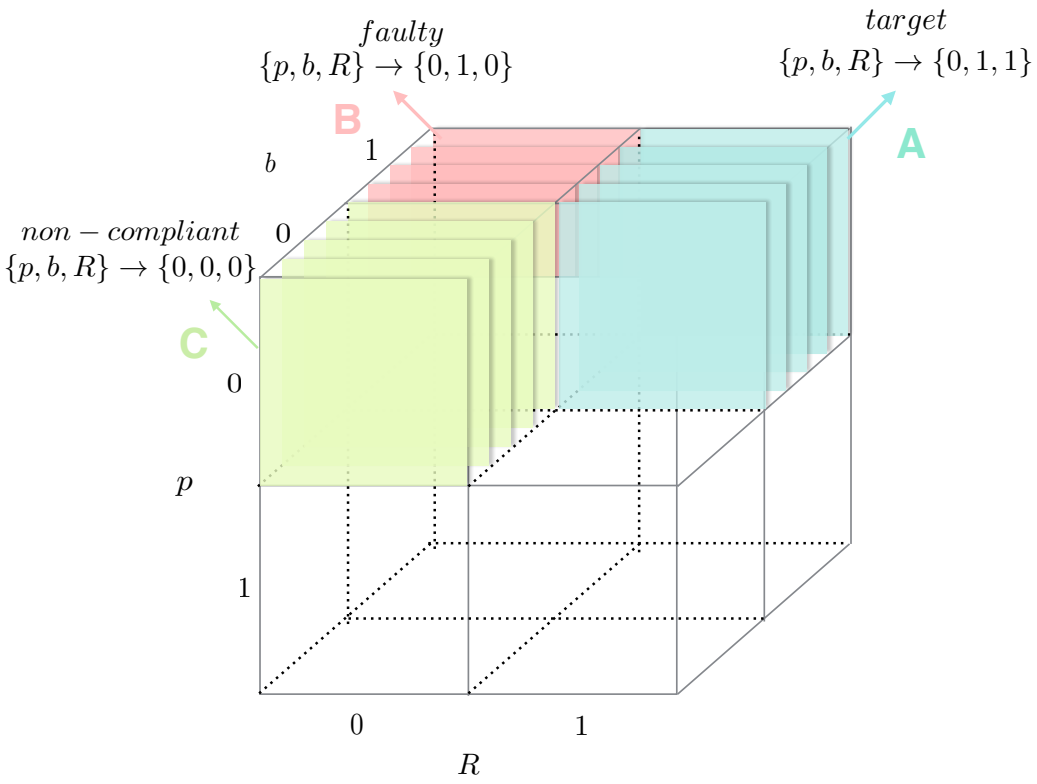
Figure 4.1: A Tree Outlining All Possible Prediction (p) and Reality (R) Combinations



The letter 'A' indicates the target situation, where an individual is predicted to have poor sleep quality ($p_t = \{0\}$), but ends up having good sleep quality ($R = \{1\}$). Letters 'B' and 'C' indicate the situation where despite the early prediction of poor quality sleep, the user still resulted in having poor sleep. This could occur for two different reasons: (*i*) the user was *non-compliant*, in that they did not follow any recommendations (indicated by C), or (*ii*) the recommendation itself was *faulty* (indicated by B). Note that the recommendation would be provided at time T_t.

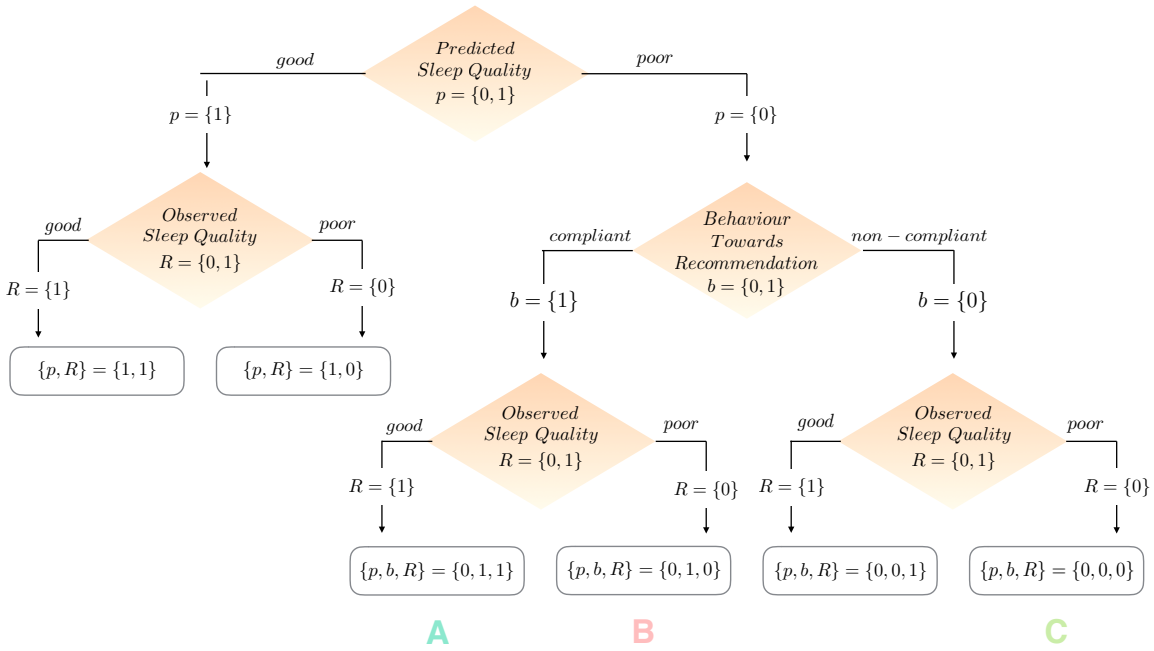
Figure 4.2 shows a 3-dimensional visualisation with the dimensions of predicted sleep quality (p), behaviour towards recommendation (b), and the resulting sleep quality (R). Figure 4.3 shows the same concept but in a decision tree format. Predicted sleep quality can be good or poor, ($p_t = \{0, 1\}$). The behaviour towards the recommendation is either compliant ($b = \{1\}$) and the user followed the recommendation, or the user was non-compliant ($b = \{0\}$) and did not follow the recommendation. Lastly, the resulting sleep quality, can also be good or poor, ($R = \{0, 1\}$). In a retrospective evaluation, determining the ratios of recommendation compliance and observed good sleep can lead to insights into the effectiveness of the recommendation system and the recommendations themselves.

Figure 4.2: A 3D visualisation of (*i*) Predicted Sleep Quality (p), (*ii*) Behaviour upon Recommendation (b), and (*ii*) Real Sleep Quality



The details on how the recommendations are determined, are described in section 4.2. The most effective and efficient way to evaluate this methodology is to do a retrospective analysis. This allows for conclusions to be drawn before a recommendation system is developed and deployed. In order to retrospectively evaluate the effectiveness of such a system, the resulting sleep quality of individuals who have compliant versus non-compliant behaviour towards recommendations should be compared. This can validate the usefulness of a recommendation, as well as the likelihood of adherence to the recommendation.

Figure 4.3: A tree visualisation of (*i*) Predicted Sleep Quality (p), (*ii*) Behaviour upon Recommendation (b), and (*ii*) Real Sleep Quality



The implementation of this evaluation calls for the comparison of time series that have similar behaviour between T_0 and T_{t-1}, i.e. x_t ,but different behaviour from T_t to T_{100}, i.e. x_{100}. This is done by first clustering the activity time series, x_t, at a time T_t. Thus all time series belonging to the same cluster, have exhibited similar behaviours at time T_t. Next is a re-clustering at time T_{100} to conceive sub-clusters. Now all time series within a given sub-cluster would exhibit similar behaviour between time T_0 and T_{100}. Note that neighbouring sub-clusters belonging to the same outer cluster are either individuals following different recommendations, or adhering versus not adhering to the recommendations. The centroids of these sub-clusters with s minimum threshold of good-to-bad sleep quality results, make up the possible recommendations, or the *behavioural recipes* to achieve

good sleep. Further details on this process are included below.

4.2 Methodology

This methodology creates a dynamic activity recommendation system, and includes the necessary steps for a retrospective analysis of the potential effectiveness of those recommendations.

4.2.1 Pre-Processing

The pre-processing stage can be broken down further into 2 stages: automated actigraphy and an intermediary implementation of RAHAR.

Automated Actigraphy

The first step of pre-processing, is to run automated actigraphy on the accelerometer output. The sleep period and the awake time need to be identified to understand an individual's behaviour. Moreover, to compute the sleep efficiency, which is critical for determining what the recommendations are, as well as which recommendation to give. See section 2.2 for further details on automated actigraphy.

Intermediary Implementation of RAHAR

To compare the different behaviours quantitatively in terms of exertion, RAHAR can be used. RAHAR calculates the accumulated proportion of time spent in each exertion level (sedentary, light, moderate, vigorous), and can be computed at any intermediary time t. Further details on this tool are defined in chapter 3.

Let x_t represent the accumulated proportion of time spent in each of the activity levels defined in chapter 3, from time 0 to time t. In other words,

$$x_t = \begin{bmatrix} S_t \\ L_t \\ M_t \\ V_t \end{bmatrix} = \left\langle \begin{bmatrix} S_0 \\ L_0 \\ M_0 \\ V_0 \end{bmatrix}, \begin{bmatrix} S_1 \\ L_1 \\ M_1 \\ V_1 \end{bmatrix}, \cdots, \begin{bmatrix} S_i \\ L_i \\ M_i \\ V_i \end{bmatrix}, \cdots, \begin{bmatrix} S_{t-1} \\ L_{t-1} \\ M_{t-1} \\ V_{t-1} \end{bmatrix}, \begin{bmatrix} S_t \\ L_t \\ M_t \\ V_t \end{bmatrix} \right\rangle$$

i.e. x_t is the output of RAHAR at time t for sedentary, light, moderate, and vigorous activity, respectively. Note, The value of t will change continuously, and the RAHAR output will as well. This will lead to the recommendations dynamically updating.

4.2.2 Clustering

For clustering, K-Means is used to cluster on the accumulated proportions, x_t. However, first the number of clusters, k, must be selected. Here, the Calinski-Harabasz index (CH), also known as the variance ratio criterion, is used [139]. This method could also be substituted for other techniques as well. The test is done on one to ten clusters and iterated 1000 times. Equation 4.1, 4.2 and 4.3 define the criterion.

$$CH = \frac{BGSS/(K-1)}{WGSS/(N-K)} = \frac{N-K}{K-1} * \frac{BGSS}{WGSS} \tag{4.1}$$

BGSS is the *between-group dispersion*, i.e. the weighted sum of the squared distances between cluster centroids, $G^{\{k\}}$ and the overall centre, G.

$$BGSS = \sum_{k=1}^{K} n^k ||G^{\{k\}} - G||^2 \tag{4.2}$$

$WGSS^{\{k\}}$ is the *within cluster dispersion* for cluster k. This summed over all the clusters is the WGSS, the *pooled within-cluster sum of squares.* M refers to the overall data matrix.

$$\begin{aligned} WGSS^{\{k\}} &= \sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2 \\ &= \frac{1}{n^k} \sum_{i<j \in I_k} ||M_i^{\{k\}} - M_j^{\{k\}}||^2 \end{aligned} \tag{4.3}$$

$$WGSS = \sum_{k=0}^{K} WGSS^{\{k\}}$$

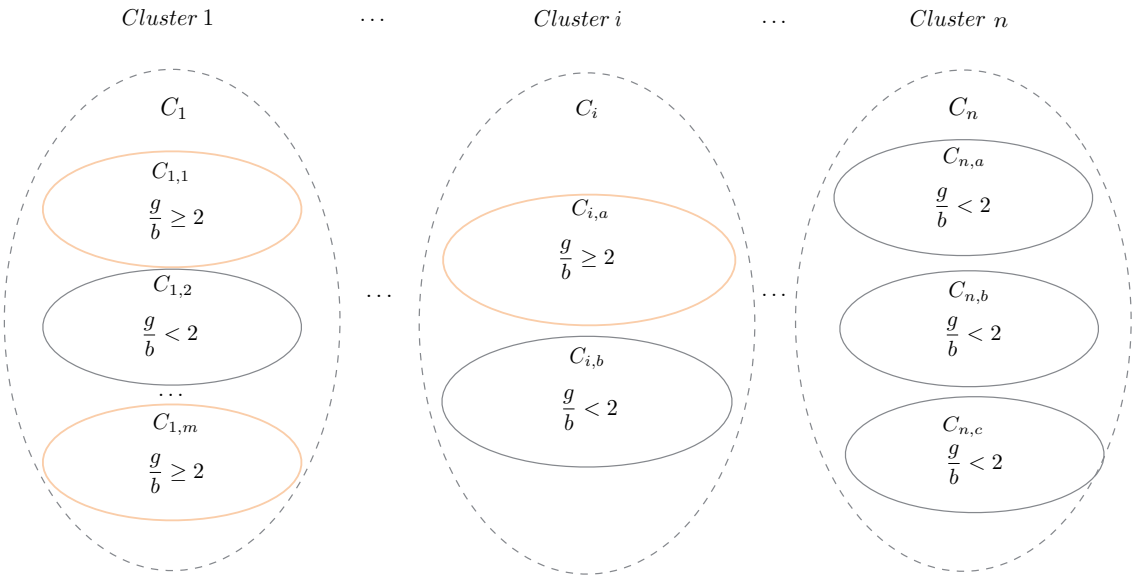
The number of clusters, k, with the highest CH, is chosen. Further details on the Calinski-Harabasz index and its implementation, can be found in the original paper [139]. Once the number of clusters, k has been determined k-means [140] is applied to all time series.

4.2.3 Sub-Clustering

In section 4.2.2, the first outer level of clusters was created based on x_t. This process is repeated within each cluster, however this time using x_{100}. In the case of retrospective analysis, x_{100} it simply represents the remainder of a person's activity until the start of the sleep period, i.e. 100% of their activity signal. In real-time recommendations x_100 represents x_j where $t < j$.

As a result, the outer clusters group together activity signals with similar behaviour between T_0 and T_{t-1}, and the sub-clusters group activity signals with similar behaviour between T_0 and T_{t-1} *and* similar behaviour between T_t and T_100 or T_j

Figure 4.4: A Visualization of the sub-clustering. The orange sub-clusters denote those whose centroids would be behavioural recipes.



4.2.4 Prediction Modelling

Before identifying the behavioural recipes, the time series need to be identified as resulting in good or poor sleep. This is determined in the pre-processing step with automated actigraphy. However, the *expected* sleep quality at T_t is also needed. If a person is expected to have poor sleep quality, the recommendation would be triggered. To determine the *expected* sleep quality, any of the previously defined modelling techniques mentioned in chapters 2 or 3 can be used.

4.2.5 Identify Behavioural Recipes

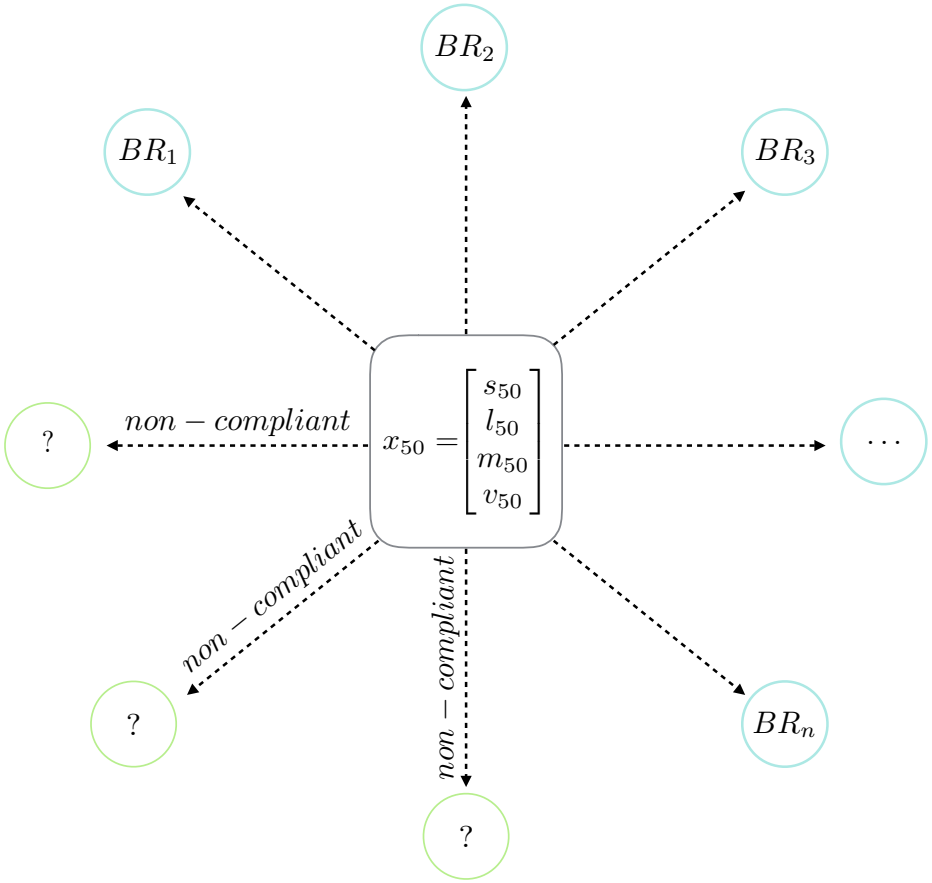
To determine which of potential sub-cluster centroids can be used as a behavioural recipe, BR_i, the ratio of activity signals that result in good sleep relative to activity signals that result in bad sleep, is computed.

The sub-clusters that have a minimum good-to-poor ratio of or above a threshold of 2, are considered behavioural recipes. This threshold could also be modified, depending on the training population or the fidelity needs of the application area.

4.2.6 Assign Recommendation

The last and final step is how to determine what recommendation to give to which user. This is computed by measuring the multi-dimensional euclidean distance between some x_t and all $BR_{i \to k}$. The centroid with the shortest distance to x_t, is the chosen recommendation. As a user's behaviour is updated throughout the day, the distance simply needs to be recalculated and the selected recommendation dynamically updated. Note that the recommendations are all proportions of time spent in different activities so it would never be the case where a recommendation would imply some behaviour "undone".

Figure 4.5: A Visualization of a user's possible reactions towards a recommendation.



4.3 Experimental Design

In this section, the dynamic activity recommendation system design outlined above is implemented and evaluated. The same clinical trial actigraphy data used throughout each chapter in this body of work, is used again.

4.3.1 Pre-Processing

The same data described in section is used. We partition the data into 75%-15%-15% random samples, for training, testing and validation, respectively. Using the raw accelerometer output, automated actigraphy is performed to identify the relevant sleep metrics including the sleep period, sleep efficiency and sleep quality.

To run RAHAR on an intermediary time point in the dataset, we choose t as 50, where the 50 denotes 50% of the length of the time series. In other words:

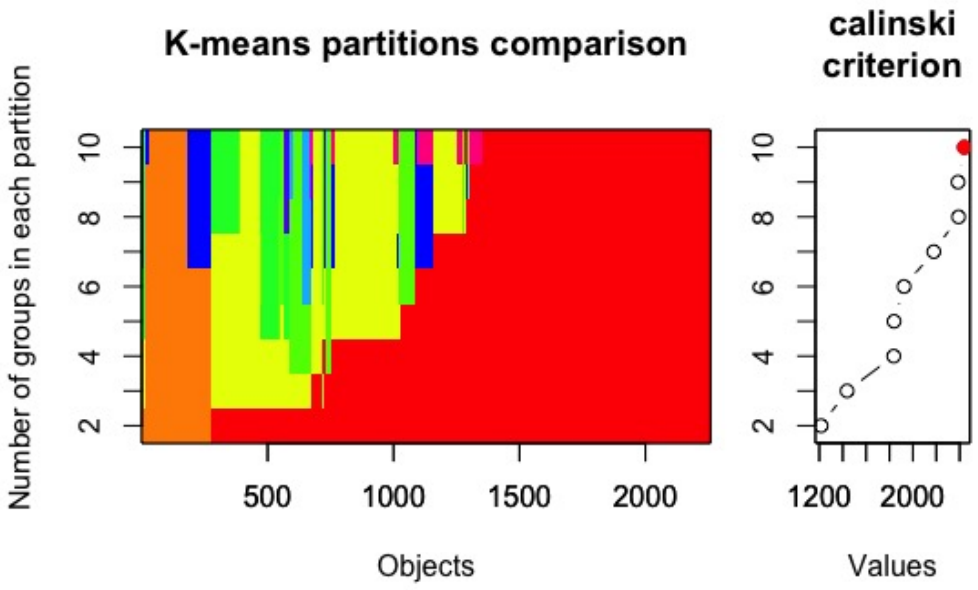
$$x_{50} = \begin{bmatrix} S_{50} \\ L_{50} \\ M_{50} \\ V_{50} \end{bmatrix} = \left\langle \begin{bmatrix} S_0 \\ L_0 \\ M_0 \\ V_0 \end{bmatrix}, \begin{bmatrix} S_1 \\ L_1 \\ M_1 \\ V_1 \end{bmatrix}, \cdots, \begin{bmatrix} S_{||T_j/2||} \\ L_{||T_j/2||} \\ M_{||T_j/2||} \\ V_{||T_j/2||} \end{bmatrix} \right\rangle$$

, where j is the length of the whole time series. Since each user has a different length time series, i.e. stays awake for a different amount of time, using a proportion of the awake time is more appropriate than setting a constant threshold.

4.3.2 Clustering

As mentioned above, the Calinski-Harabasz Index is used to determine the number of clusters to use in k-means. Figure 4.6 shows a visualisation of the criterion. The highest value indicates the best number of clusters. For simplicities purpose, the maximum number of clusters is set to ten.

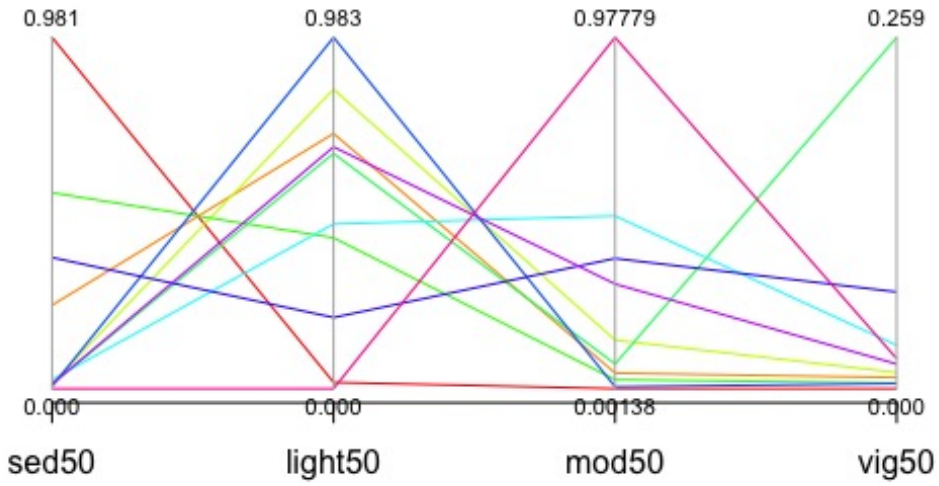
Figure 4.6: A Visualization of the Calinski-Harabasz Index used to determine the ideal number of clusters



Based on the results seen in figure 4.6, it can be concluded that k is set to 10. Now that the number of clusters to use is determined, k-means is run. To double check the validity of the value selected for k, a cluster visualisation can be assessed. Figure 4.7 is a parallel coordinates clustering visualisation, illustrating the ten clusters. The four dimensions are 'sed50', 'light50', 'mod50' and 'vig50', representing the sedentary, light, moderate and vigorous accumulated activity proportions at the point in the time series that is 50% of its total duration. The ten coloured lines zigzagging between the vertical axes represent each individual cluster. The point at which each coloured line intersects with the vertical lines indicates the value of its centroid in that dimension.

An ideal clustering illustrated in parallel coordinates will show distinctly unique lines representing each cluster. As can be seen in figure 4.7 this is the case. Although the majority of clusters seem to

Figure 4.7: A Visualization of the Clustering in Parallel Coordinates with 10 Clusters

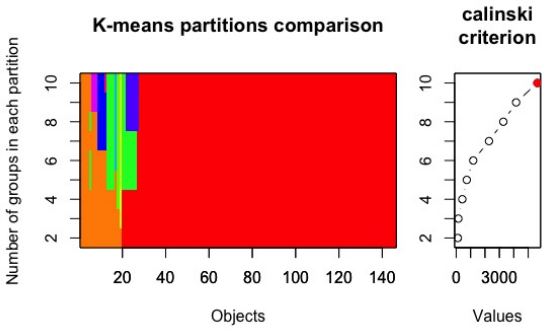


converge to a low value on the 'vig50' axis, this is representative of the data, where the majority of individuals did not engage in vigorous activity.

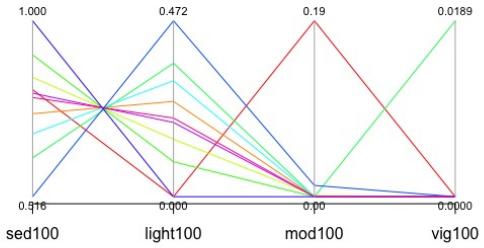
4.3.3 Sub-Clustering

Within each of the 10 clusters, the same process is iterated: k is determined from the Calinski-Harabasz Index, and a visualisation of the cluster is completed using parallel coordinates. Figure 4.8 shows the graphs for each of the ten clusters.

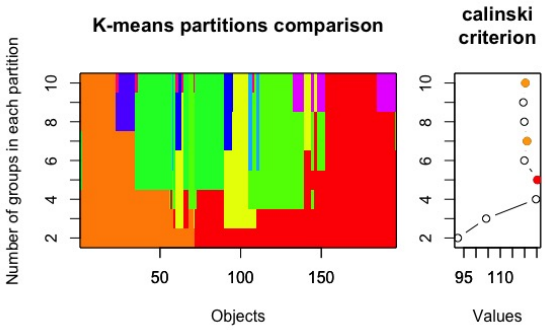
Figure 4.8: Visualisation of the subcluster size selection using the Calinski Criterion , and the cluster centroids using Parallel Coordinates



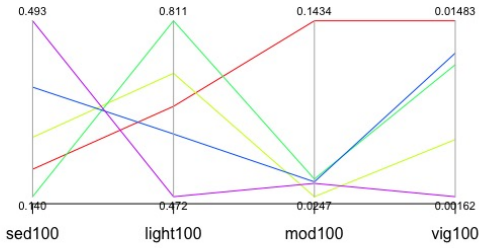
(a) Cluster 1: Calinski Criterion



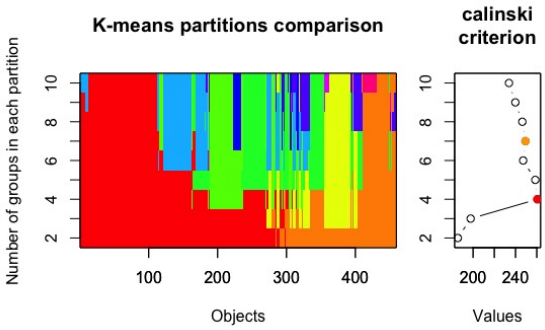
(b) Cluster 1: Parallel Coordinates



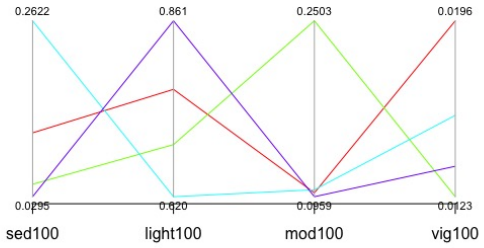
(c) Cluster 2: Calinski Criterion



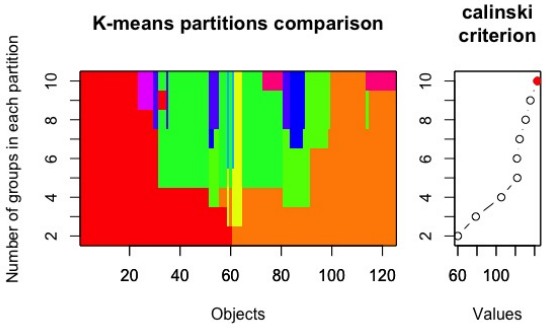
(d) Cluster 2: Parallel Coordinates



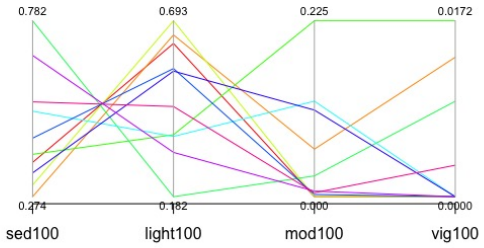
(e) Cluster 3: Calinski Criterion



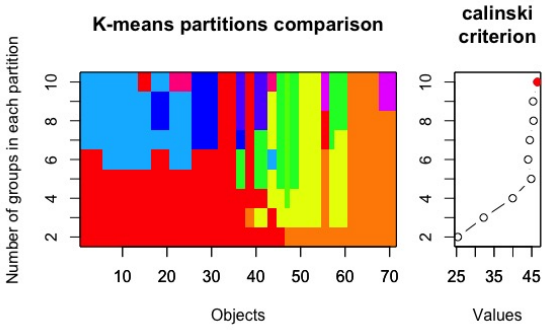
(f) Cluster 3: Parallel Coordinates



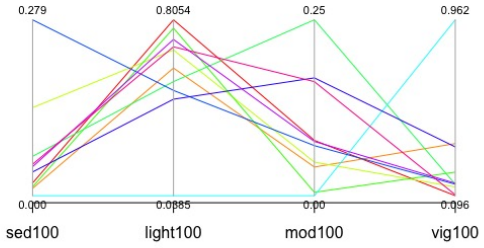
(g) Cluster 4: Calinski Criterion



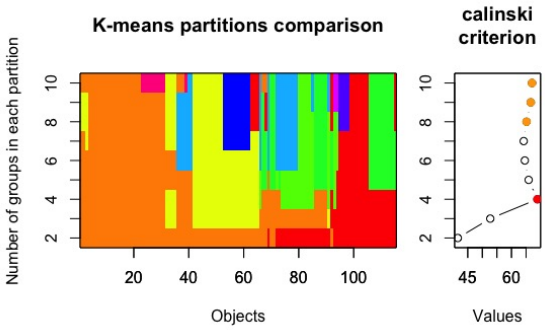
(h) Cluster 4: Parallel Coordinates



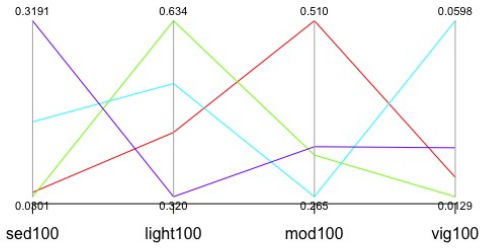
(i) Cluster 5: Calinski Criterion



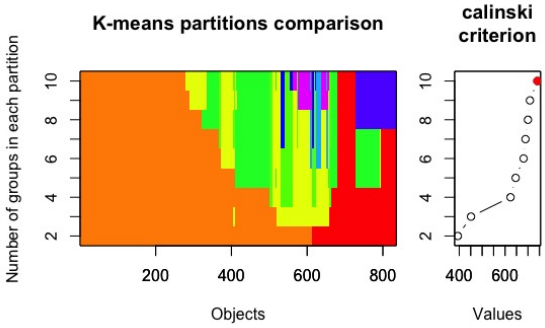
(j) Cluster 5: Parallel Coordinates



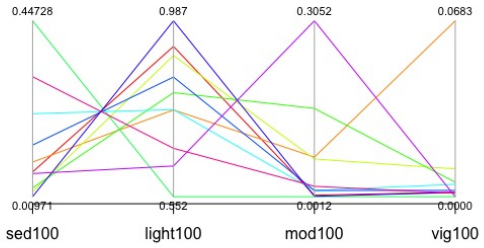
(k) Cluster 6: Calinski Criterion



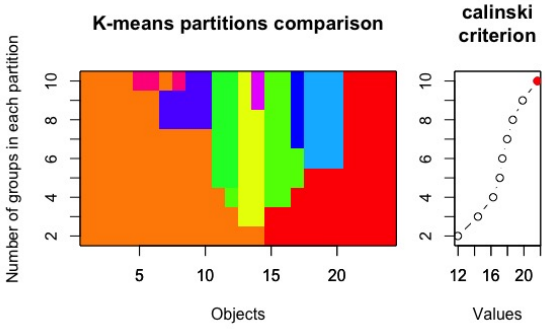
(l) Cluster 6: Parallel Coordinates



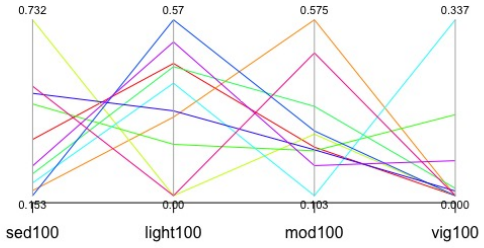
(m) Cluster 7: Calinski Criterion



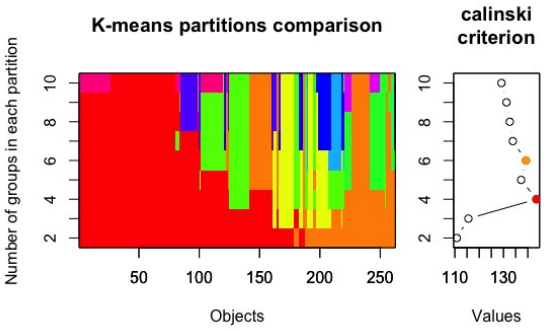
(n) Cluster 7: Parallel Coordinates



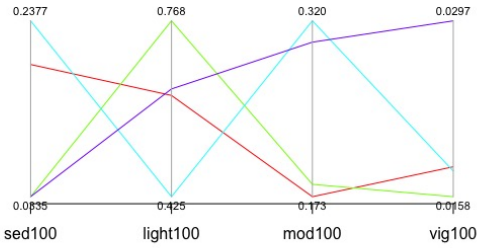
(o) Cluster 8: Calinski Criterion



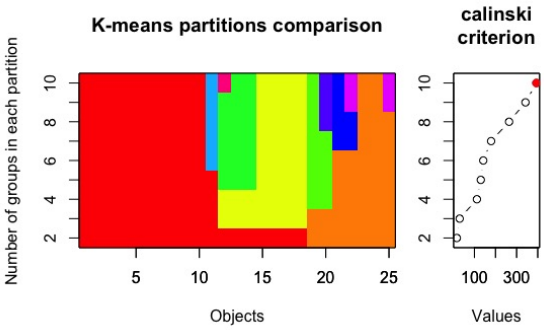
(p) Cluster 8: Parallel Coordinates



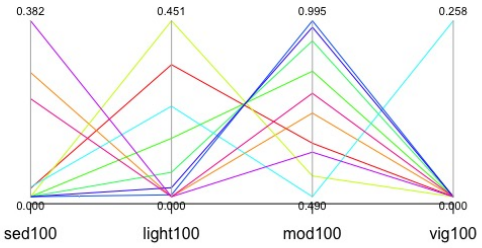
(q) Cluster 9: Calinski Criterion



(r) Cluster 9: Parallel Coordinates



(s) Cluster 10: Calinski Criterion



(t) Cluster 10: Parallel Coordinates

Table 4.1: Optimal k for each cluster based on the Calinski-Criterion

Cluster	k
1	10
2	5
3	4
4	10
5	10
6	4
7	10
8	10
9	4
10	10
Total	74

Table 4.1 shows a summary of the optimal k used to create the sub-clusters within each of the ten clusters. Thus the total number of sub-clusters for this dataset, is 74.

4.3.4 Prediction Modelling

Any high fidelity prediction algorithm can be used to predict the sleep quality at T_{50}. For simplicities sake, random forest is used here. Note that all individuals predicted to have poor sleep would receive a recommendation, i.e. 869 of the 3226 individuals.

4.3.5 Identify Behavioural Recipes

At this point, the sub-clusters have been defined, as well as the *predicted* and *observed* sleep quality for each time series. Thus the ratio of the observed, or real (R), good to bad sleep ratio can be

computed for each sub-cluster. Tables 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 and 4.11 show the proportions and ratios for each of the corresponding sub-clusters. The centroid of any sub-cluster with a ratio greater than a threshold of 2, is considered a behavioural recipe.

Table 4.2: Good/Bad sleep ratios for each sub-cluster in Cluster 1

Cluster	1	1	1	1	1	1	1	1	1	1
Sub-Cluster	1	2	3	4	5	6	7	8	9	10
Prop. of Poor Quality	1.00	1.00	0.80	1.00	0.30	0.50	0.40	0.69	0.50	1.00
Prop. of Good Quality	0.00	0.00	0.20	0.00	0.70	0.50	0.60	0.31	0.50	0.00
Ratio	0.00	0.00	0.25	0.00	2.33	0.50	1.50	0.45	1.00	0.00

Table 4.3: Good/Bad sleep ratios for each sub-cluster in Cluster 2

Cluster	2	2	2	2	2
Sub-Cluster	1	2	3	4	5
Prop. of Poor Quality	0.61	0.61	0.45	0.72	0.81
Prop. of Good Quality	0.39	0.39	0.55	0.28	0.19
Ratio	0.04	0.04	1.22	0.39	0.24

Table 4.4: Good/Bad sleep ratios for each sub-cluster in Cluster 3

Cluster	3	3	3	3
Sub-Cluster	1	2	3	4
Prop. of Poor Quality	0.58	0.34	0.64	0.28
Prop. of Good Quality	0.42	0.66	0.36	0.72
Ratio	0.72	1.91	0.56	2.62

Table 4.5: Good/Bad sleep ratios for each sub-cluster in Cluster 4

Cluster	4	4	4	4	4	4	4	4	4	4
Sub-Cluster	1	2	3	4	5	6	7	8	9	10
Prop. of Poor Quality	0.17	0.25	0.22	1.00	0.75	0.53	1.00	0.60	0.83	0.63
Prop. of Good Quality	0.83	0.75	0.78	0.00	0.25	0.47	0.40	0.62	0.17	0.37
Ratio	5.00	3.00	3.50	1.33	0.00	0.33	0.89	0.67	0.20	0.60

Table 4.6: Good/Bad sleep ratios for each sub-cluster in Cluster 5

Cluster	5	5	5	5	5	5	5	5	5	5
Sub-Cluster	1	2	3	4	5	6	7	8	9	10
Prop. of Poor Quality	0.33	0.57	0.33	0.00	0.60	0.00	0.92	0.50	0.18	0.00
Prop. of Good Quality	0.67	0.43	0.67	1.00	0.40	1.00	0.08	0.50	0.81	1.00
Ratio	2.00	0.75	2.00	∞	0.67	∞	0.08	1.00	4.50	7.00

Table 4.7: Good/Bad sleep ratios for each sub-cluster in Cluster 6

Cluster	6	6	6	6
Sub-Cluster	1	2	3	4
Prop. of Poor Quality	0.26	0.41	0.56	0.69
Prop. of Good Quality	0.74	0.59	0.44	0.31
Ratio	2.86	1.42	0.80	0.44

Table 4.8: Good/Bad sleep ratios for each sub-cluster in Cluster 7

Cluster	7	7	7	7	7	7	7	7	7	7
Sub-Cluster	1	2	3	4	5	6	7	8	9	10
Prop. of Poor Quality	0.44	0.66	0.21	0.33	0.60	0.59	0.49	0.15	0.31	0.66
Prop. of Good Quality	0.56	0.34	0.79	0.67	0.40	0.41	0.51	0.85	0.69	0.34
Ratio	1.28	0.52	3.82	2.00	0.67	0.70	1.04	5.69	2.20	0.53

Table 4.9: Good/Bad sleep ratios for each sub-cluster in Cluster 8

Cluster	8	8	8	8	8	8	8	8	8	8
Sub-Cluster	1	2	3	4	5	6	7	8	9	10
Prop. of Poor Quality	0.50	0.00	1.00	0.67	0.25	0.50	0.50	0.50	0.33	0.50
Prop. of Good Quality	0.50	1.00	0.00	0.33	0.75	0.50	0.50	0.50	0.67	0.50
Ratio	1.00	∞	0.00	0.50	3.00	1.00	1.00	1.00	2.00	1.00

Table 4.10: Good/Bad sleep ratios for each sub-cluster in Cluster 9

Cluster	9	9	9	9
Sub-Cluster	1	2	3	4
Prop. of Poor Quality	0.49	0.25	0.70	0.41
Prop. of Good Quality	0.51	0.75	0.30	0.59
Ratio	1.04	3.03	0.44	1.46

Table 4.11: Good/Bad sleep ratios for each sub-cluster in Cluster 10

Cluster	10	10	10	10	10	10	10	10	10	10
Sub-Cluster	1	2	3	4	5	6	7	8	9	10
Prop. of Poor Quality	0.33	0.00	0.50	0.00	0.00	0.50	0.50	0.00	0.50	0.00
Prop. of Good Quality	0.67	1.00	0.50	1.00	1.00	0.50	0.50	1.00	0.50	1.00
Ratio	2.00	∞	1.00	∞	∞	1.00	1.00	5.00	1.00	∞

4.3.6 Determine Recommendation

For real-time recommendations, the multidimensional euclidean distance from an individual's x_t to all of the centroids included in table 4.12 is computed. The shortest distance indicates the chosen recommendation for suggestion.

From the above table it is clear that not all clusters contain a behavioural recipe. This is an issue for a retrospective evaluation where it is necessary to do a comparison on the resulting sleep quality of people who have the same predicted sleep quality (p_{50}), but either do or do not follow the recommendation. Thus any time series in a cluster that does not contain a behavioural recipe is reassigned. The reassignment is determined by measuring the distance from x_t to each of the centroids. This is the same process as the recommendation decision, and essentially is determining what recommendation would be given to that individual.

Table 4.12: The Behavioural Recipes

Cluster	Sub-cluster	Sedentary	Light	Moderate	Vigorous
3	4	0.029	0.861	0.096	0.014
4	1	0.373	0.627	0.000	0.000
4	2	0.274	0.652	0.061	0.013
4	3	0.307	0.693	0.000	0.000
5	1	0.020	0.805	0.079	0.096
5	3	0.140	0.674	0.048	0.138
5	4	0.014	0.769	0.005	0.212
5	9	0.046	0.720	0.077	0.157
5	10	0.050	0.688	0.162	0.100
6	1	0.037	0.435	0.510	0.018
7	3	0.022	0.901	0.066	0.011
7	4	0.032	0.809	0.154	0.005
7	8	0.010	0.987	0.002	0.001
7	9	0.067	0.628	0.305	0.000
9	2	0.034	0.768	0.183	0.015

4.3.7 Evaluate Recommendation

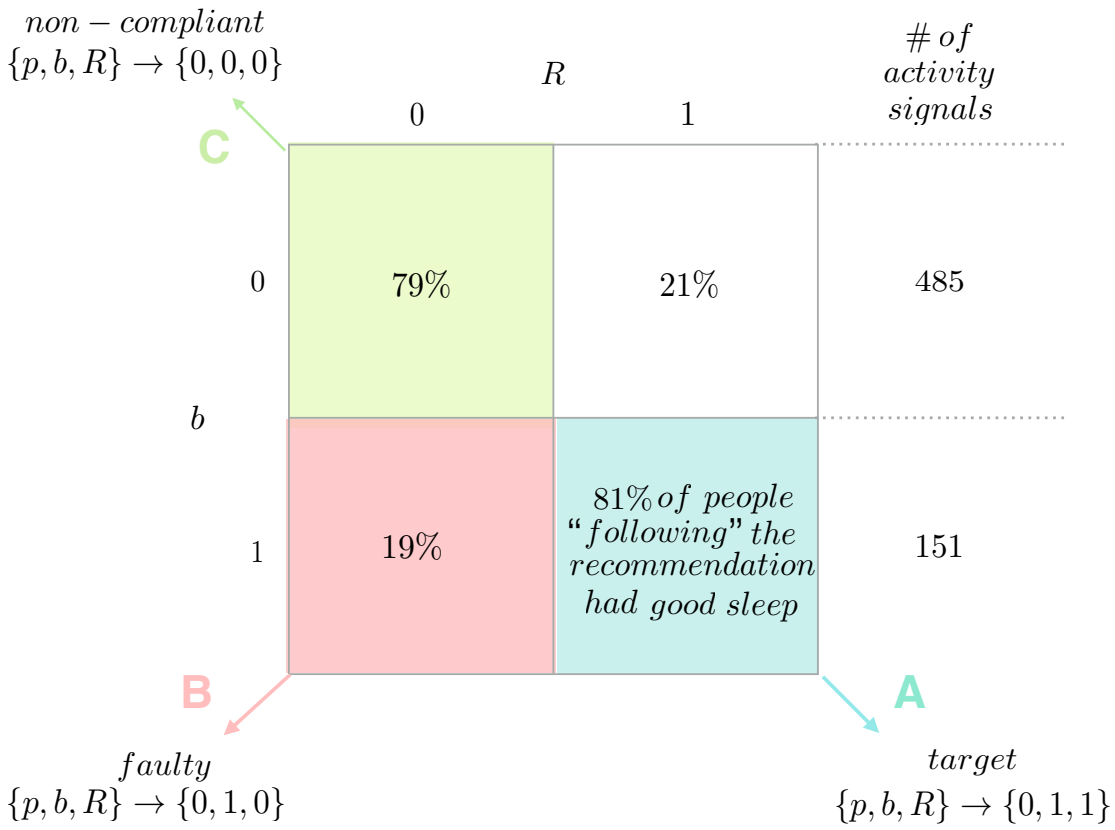
In order to evaluate whether a recommendation was taken, a comparison needs to be done on the resulting sleep quality of people who have the same predicted sleep quality (p_{50}), but either (*i*) follow the recommendation ($b = \{1\}$), or (*ii*) do not ($b = \{0\}$).

Two activity signals belonging to the same cluster, but different sub-clusters, indicates that the signals followed similar behaviours between T_0 and T_{50}, and different behaviours between T_{50} and

T_{100}. Assume both of these activity signals were also predicted to have poor sleep, i.e. they received a recommendation. The signals that belong to a sub-cluster with a good-to-bad ratio greater than 2, are in the same cluster as a behaviour recipe. This suggests that they exhibit similar behaviour to the recommendation, so it naturally concludes that they followed, or adhered, to the recommendation. Conversely, it can be concluded that any signals belonging to sub-clusters with a good-to-bad ratio below 2, did not adhere to the recommendation, and were non-compliant. Figure 4.5 illustrates the possible behaviours towards a recommendation.

Figure 4.6 contains summary results from the experiment. Note that in this dataset the majority of activity signals are predicted to have good sleep at T_{50}, and do indeed have good sleep. This is one of the causes of such low proportions visible in the confusion matrix.

Figure 4.6: A Matrix with Results from a Retrospective Analysis.



Chapter 5

Conclusion

The overall conjecture is that our phenome, (i.e. our manifestation of self), is a function of (*i*) our genome, the fundamental DNA that makes up who we are and, (*ii*) our exposome, the entire set of exposures that we endure during our lifetime. While we cannot alter our genetics, at least without extreme measures, we can alter our expotype by our everyday behaviour. The genome has been mapped by the human genome project. Since then, there has been a lot of research in computational biology, working on drawing insights and interpreting this DNA. Thanks to the internet of things and quantified self movement, sensors are ubiquitous, and we now have the tools to also collect information and quantify the human exposome.

Our exposome consists of our living arrangements, environment, air quality, employment status, disease exposure, digital exposure, and more. One additional component is sleep. There have been studies upon studies that have found good sleep to be correlated with health and happiness, and even more so that poor sleep can have serious side effects. Wearable devices can be a catalyst for innovative research by quantifying human behaviours and activities, and determining the effect they have on sleep and ultimately quality of life.

This body of work makes an effort to migrate the traditional and cumbersome current sleep science process to a computational data-driven approach. Thus there are contributions both to the field of sleep science, as well as computer science.

5.1 Sleep Science Contributions

Sleep problems range from serious conditions such as sleep apnea and insomnia, to minor day-to-day discomfort and fatigue. As a result the sleep science process can range from tools for consumer wearables to clinical care. With large scale projects such as the All of Us research project, there is also a need for tools to be used in the research pipeline.

5.1.1 Consumer Tools

There has been a societal transition in healthcare, moving the point of care from mainframe institutions to the home and community. Consumers want to engage in their own welfare. One of the biggest impediments to participatory care, is the overwhelming influx of data. Wearable devices provide a huge amount of big data, and are particularly useful for activity and sleep tracking. Translating this data into real world impact for the individual is critical.

In chapter 2, a new deep learning architecture was designed, specifically with wearable devices in mind. The TB-LSTM architecture provides a high fidelity screening mechanism that can be embedded within any wearable device. If a mobile application with this algorithm were created to pull and model the raw data from a wearable device, users would instantly have the ability to better understand the relationship between their physical activity and sleep quality, and moreover, could change their behaviour. In fact, if Fitbit, Nokia, Apple or Samsung, were to integrate this prediction algorithm into their popular wearable devices, this screening test could have profound reach, and potentially reduce the number of undiagnosed sleep problems worldwide.

Chapter 4 introduces a dynamic activity recommendation system. This system could also be integrated into mobile applications or wearable devices. The implications of integrating this system can lead to consumers developing a deeper understanding of their own personal sleep patterns, and lead to proactive involvement in their own sleep health. Moreover, the recommendations provide concrete actionable steps towards behaviour change. A successful recommendation engine for sleep quality has the potential to reduce fatigue in consumers, patients, and society as a whole.

5.1.2 Clinical Support

From a clinical support perspective, adding a screening stage to the sleep science pipeline can allow for more targeted care. The current sleep science process doesn't include any mechanism for large scale screening and thus a large number of chronic sleep conditions remain undiagnosed. Embedding a screening mechanism into consumer devices, can notify potential patients in need of clinical care, to take action. Moreover, consumers with minor sleep difficulties could be handled outside the clinic, reducing the overall bottleneck of sleep clinic visits.

Chapter 3 described a robust and automated human activity recognition algorithm called RA-HAR. This algorithm advances automated actigraphy by additionally automating the annotation of activity exertion levels. This tool reduces the overhead of manual examination of actigraphy output. Moreover, it extends current actigraphy capabilities by providing high-level insights into a person's activity. By providing deeper insights, clinicians can break away from the trial-and-error iterative cycle of *standard of care*.

5.1.3 Research Pipeline

Large cohort clinical trials collect an abundance of wearable device data, and due to the cumbersome analysis process, often lie untouched. With projects like the Kavli HUMAN project and the All of Us research initiative, it is never been more vital to streamline the research pipeline. Automated

actigraphy and RAHAR are tools that can interpret and clean raw wearable device data and return it in a palatable form ready for research.

5.2 Computer Science Contributions

Each chapter in this body of work contains a key methodology that advances the field of computer science for wearable devices.

5.2.1 Time-Batched LSTM

The TB-LSTM architecture is a new deep learning architecture targeting the modelling of wearable device data. Although recurrent neural networks are esteemed for their abilities in managing time series data, the results in chapter 2 show that recurrent neural networks were unable to model wearable device data effectively. Both simple RNNs and LSTM RNNs operate on sequences, where each time step comprises only one activity value. This often results in very long sequences. As a consequence, RNNs cannot compose higher-level features effectively due to the low-dimensional input at each time step. They also suffer from vanishing gradient problems due to the disproportionate length of the sequences. The TB-LSTM architecture proved its ability to succeed where past methods have fallen short.

5.2.2 Human Activity Recognition

Chapter 3 defined a unique human activity recognition algorithm that focused on personal exertion levels rather than activity type. While RAHAR was tested specifically on its ability to analyse sleep data, it could be utilized in any other domain areas as well. One of the principal progressions of RAHAR is the personalised nature of its computation. Change point detection divides the time series into exertion phases based solely on an individual's own time series. This aligns with the

notion that the amount of effort an individual spends on a task, may differ greatly from another individual's exertion for the same task. This makes RAHAR a unique and innovative solution to analysing wearable devices.

5.2.3 Recommendation System for Behavioural Change with a Retrospective Evaluation

The dynamic activity recommendation system outlined in chapter **??** is a unique approach to activity recommendation. The methodology evaluates the system before it is deployed. Doing a retrospective analysis like so, can allow for quick improvements. Moreover, it is a great tool to explore clinical trial data or projects like All of Us, where intervention is not appropriate. Because the recommendation system is based off of RAHAR, it provides personalised insights. Lastly, the recommendation can be computed in constant time, thus real-time, and dynamically updated, as an individual's behaviour changes throughout the day.

5.3 Future Work

While this thesis provides a number of novel approaches, there are many more ways to advance the field of computational sleep science and create new techniques to analyse wearable devices.

5.3.1 Systems Development

The true impact of the work described in this thesis can be realized by developing these tools into a software suite. There are a multitude of tasks that automated actigraphy and RAHAR simplify. Building out a fully functional system that integrates these tools into the actigraphy workflow, stands to make an instant impact to the sleep scientist community. Clinicians and behavioural scientists could use these tools for the evaluation of patients or clinical trial cohorts alike.

Furthermore, developing the recommendation system described in chapter 4 would create the first activity recommendation system targeting sleep quality, of its kind. Integrating this system with a wearable device would also create instant impact empowering users to take control rather than relying on a serendipitous night of good sleep.

5.3.2 Continued Analysis

Chapter 4 took an exploratory look at activity recommendations for sleep quality. This research opens the doors for much further analysis. One of the interesting concepts is to run the clustering and prediction modelling, on one individual's data collected over time. The recommendations would be based on a user's personal sleep history and are likely to be even more effective.

One of the key components in this body of work has been prediction modelling algorithms for sleep quality. Prediction algorithm research is an ever-evolving field. New techniques such as reinforcement learning show promise, and could be particularly good at modelling human behaviour given that the inspiration behind the algorithms lies in behavioural psychology.

5.3.3 Data

Lastly, this thesis starts the transition of sleep science into a computational, data-driven field. The majority of this work focused on wearable device data. Particularly for screening, it was important to use minimal sensor input, so that the algorithms would perform on almost any wearable device. However, integrating other sensors, such as heart rate monitors, electrodes, luminosity sensors, can lead to additional improvement.

Sleep quality is not just effected by physical activity. Integrating nutritional information, could provide additional insights, particularly caffeine consumption. Electronic health records contain a wealth of information into an individual's well-being. It is known that poor sleep quality can lead to exacerbation of many diseases. It can also be a side effect of serious health issues. Furthermore,

exploring the relationship between sleep quality and cognitive load is bound to be of interest. Lastly sleep patterns can often be hereditary. Examining genomic data and its relationship to sleep quality is another are for research.

...

Bibliography

[1] "Polysomnography," http://dxline.info/diseases/polysomnography , accessed: 2017-10-29.

[2] "Home sleep test," https://www.resmed.com/us/en/commercial-partner/airsolutions/home-sleep-testing.html, accessed: 2017-10-29.

[3] W. H. Organization *et al.*, "Global health and aging," *Acedido março*, vol. 5, p. 2015, 2011.

[4] ——, "Global health observatory (gho) data," *URL. Available form: http://www. who. int/gho/tb/en*, 2015.

[5] D. E. Bloom, E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. R. Bloom, S. Fathima, A. B. Feigl, T. Gaziano, A. Hamandi, M. Mowafi *et al.*, "The global economic burden of noncommunicable diseases," Program on the Global Demography of Aging, Tech. Rep., 2012.

[6] I. Anderson and E. Jarawan, "The economic costs of noncommunicable diseases in the pacific islands," The World Bank, Tech. Rep., 2012.

[7] A. Caplin, M. Luo, and K. McGarry, "Measuring and modeling intergenerational links in relation to long-term care," *Economic Inquiry*, 2016.

[8] K. M. Langa and D. Cutler, "Opportunities for new insights on the life-course risks and outcomes of cognitive decline in the kavli human project," *Big data*, vol. 3, no. 3, pp. 189–192, 2015.

[9] F. T. Shaya and V. V. Chirikov, "Decision support tools to optimize economic outcomes for type 2 diabetes." *The American journal of managed care*, vol. 17, pp. S377–83, 2011.

[10] T. van der Molen and S. Schokker, "Primary prevention of chronic obstructive pulmonary disease in primary care," *Proceedings of the American Thoracic Society*, vol. 6, no. 8, pp. 704–706, 2009.

[11] S. Alladi, T. H. Bak, V. Duggirala, B. Surampudi, M. Shailaja, A. K. Shukla, J. R. Chaudhuri, and S. Kaul, "Bilingualism delays age at onset of dementia, independent of education and immigration status," *Neurology*, vol. 81, no. 22, pp. 1938–1944, 2013.

[12] C. H. Schenck, B. F. Boeve, and M. W. Mahowald, "Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: a 16-year update on a previously reported series," *Sleep medicine*, vol. 14, no. 8, pp. 744–748, 2013.

[13] P. M. I. W. Group *et al.*, "Report to the advisory committee to the director: The precision medicine initiative cohort program?building a research foundation for 21st century medicine," *Washington, DC: National Institutes of Health*, 2015.

[14] D. Ausiello and S. Lipnick, "Real-time assessment of wellness and disease in daily life," *Big data*, vol. 3, no. 3, pp. 203–208, 2015.

[15] O. Azmak, H. Bayer, A. Caplin, M. Chun, P. Glimcher, S. Koonin, and A. Patrinos, "Using big data to understand the human condition: the kavli human project," *Big data*, vol. 3, no. 3, pp. 173–188, 2015.

[16] L. Fernandez-Luque, M. Singh, F. Ofli, Y. A. Mejova, I. Weber, M. Aupetit, S. K. Jreige, A. Elmagarmid, J. Srivastava, and M. Ahmedna, "Implementing 360 quantified self for childhood obesity: feasibility study and experiences from a weight loss camp in qatar," *BMC medical informatics and decision making*, vol. 17, no. 1, p. 37, 2017.

[17] N. Newswire, "Are consumers really interested in wearing tech on their sleeves? march 20, 2014," 2015.

[18] T. M. Research, "Forecasted value of the global wearable devices market from 2012 to 2018," *Statista*, 2017.

[19] T. W. Strine and D. P. Chapman, "Associations of frequent sleep insufficiency with health-related quality of life and health behaviors," *Sleep medicine*, vol. 6, no. 1, pp. 23–27, 2005.

[20] H. R. Colten, B. M. Altevogt *et al.*, *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press, 2006.

[21] K. L. Knutson, A. M. Ryden, B. A. Mander, and E. Van Cauter, "Role of sleep duration and quality in the risk and severity of type 2 diabetes mellitus," *Archives of internal medicine*, vol. 166, no. 16, pp. 1768–1774, 2006.

[22] E. Kasasbeh, D. S. Chi, and G. Krishnaswamy, "Inflammatory aspects of sleep apnea and their cardiovascular consequences," *Southern Medical Journal-Birmingham Alabama-*, vol. 99, no. 1, p. 58, 2006.

[23] M. J. Murphy and M. J. Peterson, "Sleep disturbances in depression," *Sleep medicine clinics*, vol. 10, no. 1, pp. 17–23, 2015.

[24] A. M. Williamson and A.-M. Feyer, "Moderate sleep deprivation produces impairments in cognitive and motor performance equivalent to legally prescribed levels of alcohol intoxication," *Occupational and environmental medicine*, vol. 57, no. 10, pp. 649–655, 2000.

[25] C. M. Shapiro and W. C. Dement, "Abc of sleep disorders. impact and epidemiology of sleep disorders." *BMJ: British Medical Journal*, vol. 306, no. 6892, p. 1604, 1993.

[26] A. Sassani, L. J. Findley, M. Kryger, E. Goldlust, C. George, and T. M. Davidson, "Reducing motor-vehicle collisions, costs, and fatalities by treating obstructive sleep apnea syndrome," *SLEEP-NEW YORK THEN WESTCHESTER-*, vol. 27, no. 3, pp. 453–458, 2004.

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

[28] T. Arora and S. Taheri, "Sleep optimization and diabetes control: A review of the literature," *Diabetes Therapy*, vol. 6, no. 4, pp. 425–468, 2015.

[29] S. Abdullah, M. Matthews, E. L. Murnane, G. Gay, and T. Choudhury, "Towards circadian computing: early to bed and early to rise makes some of us unhealthy and sleep deprived," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2014, pp. 673–684.

[30] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 477–486.

[31] Y. H. Min, J. W. Lee, Y.-W. Shin, M.-W. Jo, G. Sohn, J.-H. Lee, G. Lee, K. H. Jung, J. Sung, B. S. Ko *et al.*, "Daily collection of self-reporting sleep disturbance data via a smartphone app in breast cancer patients receiving chemotherapy: a feasibility study," *Journal of medical Internet research*, vol. 16, no. 5, p. e135, 2014.

[32] A. Huffington, *The Sleep Revolution: Transforming Your Life, One Night at a Time*. Harmony, 2016.

[33] R. J. Raymann, W. N. Dougherty, N. Divya, D. M. Lambert, S. Greer, and T. R. Gruber, "Adjusting alarms based on sleep onset latency," Jun. 27 2017, uS Patent 9,692,874.

[34] L. Rabb, A. Colaco, M. Dixon, G. A. Kirmani, L. Villaran, K. L. Herman, B. James, C. M. Davis, and Y. Modi, "Using active ir sensor to monitor sleep," Nov. 5 2015, uS Patent App. 14/933,069.

[35] M.-H. Lee and S.-w. Bang, "Apparatus and/or method for inducing sound sleep and waking," Jun. 7 2011, uS Patent 7,956,755.

[36] R. Auphan, F. Dusanter, R. Yang, N. Buard, C. Hutchings, B. Rechke, and J. Gautier, "System and method to monitor and assist individual's sleep," Jul. 21 2014, uS Patent App. 14/336,856.

[37] P. Phillips, C. Heneghan, and T. Murray, "System and method for determining sleep stage," Sep. 19 2013, uS Patent App. 14/031,553.

[38] J. Park, M. M. Martinez, M. H. Berlinger, A. Ringrose, D. J. Clifton, S. E. McKinney, and G. Amit, "Wristband health tracker," Feb. 9 2016, uS Patent D749,002.

[39] U. S. D. of Health and H. Services, "National sleep foundation," 2015.

[40] A. Rechtschaffen, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Public health service*, 1968.

[41] M. Hirshkowitz, "The history of polysomnography: Tool of scientific discovery," *Sleep Medicine: A Comprehensive Guide to Its Development, Clinical Milestones, and Advances in Treatment*, pp. 91–100, 2015.

[42] N. J. Douglas, "Home diagnosis of the obstructive sleep apnoea/hypopnoea syndrome," *Sleep medicine reviews*, vol. 7, no. 1, pp. 53–59, 2003.

[43] D. A. Dean, A. L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S. S. Sahoo, C. P. Jayapandian, L. Cui, M. G. Morrical *et al.*, "Scaling up scientific discovery in sleep medicine: The national sleep research resource," 2016.

[44] R. Budhiraja, R. Thomas, M. Kim, and S. Redline, "The role of big data in the management of sleep-disordered breathing," *Sleep medicine clinics*, vol. 11, no. 2, pp. 241–255, 2016.

[45] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–469, 1992.

[46] M. Enomoto, T. Endo, K. Suenaga, N. Miura, Y. Nakano, S. Kohtoh, Y. Taguchi, S. Aritake, S. Higuchi, M. Matsuura *et al.*, "Newly developed waist actigraphy and its sleep/wake scoring algorithm," *Sleep and Biological Rhythms*, vol. 7, no. 1, pp. 17–22, 2009.

[47] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Aaai*, vol. 5, no. 2005, 2005, pp. 1541–1546.

[48] J. Baek, G. Lee, W. Park, and B.-J. Yun, "Accelerometer signal processing for user activity detection," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2004, pp. 610–617.

[49] A. Sathyanarayana, F. Ofli, L. Luque, J. Srivastava, A. Elmagarmid, T. Arora, and S. Taheri, "Robust automated human activity recognition and its application to sleep research," in *2016 IEEE International Conference on Data Mining Workshop (ICDMW)*, Dec 2016.

[50] L. Gershell, "Insomnia market," *Nature Reviews Drug Discovery*, vol. 5, no. 1, pp. 15–16, 2006.

[51] E. Wooltorton, "Herbal kava: reports of liver toxicity," *Canadian Medical Association Journal*, vol. 166, no. 6, pp. 777–777, 2002.

[52] E. A. Belongia, C. W. Hedberg, G. J. Gleich, K. E. White, A. N. Mayeno, D. A. Loegering, S. L. Dunnette, P. L. Pirie, K. L. MacDonald, and M. T. Osterholm, "An investigation of the cause of the eosinophilia–myalgia syndrome associated with tryptophan use," *New England Journal of Medicine*, vol. 323, no. 6, pp. 357–365, 1990.

[53] "Normative values of polysomnographic parameters in childhood and adolescence: Quantitative sleep parameters," *Sleep Medicine*, vol. 12, no. 6, pp. 542 – 549, 2011.

[54] R. Williams, I. Karacan, and C. Hursch, *Electroencephalography (Eeg) of Human Sleep: Clinical Applications*, ser. A Wiley biomedical-health publication.

[55] B. M. Altevogt, H. R. Colten *et al.*, *Sleep disorders and sleep deprivation: an unmet public health problem*. National Academies Press, 2006.

[56] M. A. Kredlow, M. C. Capozzoli, B. A. Hearon, A. W. Calkins, and M. W. Otto, "The effects of physical activity on sleep: a meta-analytic review," *Journal of behavioral medicine*, vol. 38, no. 3, pp. 427–449, 2015.

[57] H. S. Driver and S. R. Taylor, "Exercise and sleep," *Sleep medicine reviews*, vol. 4, no. 4, pp. 387–402, 2000.

[58] M. Chennaoui, P. J. Arnal, F. Sauvet, and D. Léger, "Sleep and exercise: a reciprocal issue?" *Sleep medicine reviews*, vol. 20, pp. 59–72, 2015.

[59] W. Wu, S. Dasgupta, E. E. Ramirez, C. Peterson, and G. J. Norman, "Classification accuracies of physical activities using smartphone motion sensors," *Journal of medical Internet research*, vol. 14, no. 5, 2012.

[60] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, and S. Taheri, "Sleep quality prediction from wearable data using deep learning," *JMIR mHealth and uHealth*, vol. 4, no. 4, 2016.

[61] O. Shmiel, T. Shmiel, Y. Dagan, and M. Teicher, "Data mining techniques for detection of sleep arousals," *Journal of neuroscience methods*, vol. 179, no. 2, pp. 331–337, 2009.

[62] M. Längkvist, L. Karlsson, and A. Loutfi, "Sleep stage classification using unsupervised feature learning," *Advances in Artificial Neural Systems*, vol. 2012, p. 5, 2012.

[63] "Actigraph gt3x+," http://actigraphcorp.com/products-showcase/activity-monitors/actigraph-wgt3x-bt/, accessed: 2017-10-29.

[64] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, no. 2, pp. 251–257, Mar. 1991.

[65] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014.

[66] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[67] T. Mikolov, *Statistical Language Models based on Neural Networks.* PhD thesis, Brno University of Technology, 2012.

[68] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[69] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[70] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[71] J. Behar, A. Roebuck, J. S. Domingos, E. Gederi, and G. D. Clifford, "A review of current sleep screening applications for smartphones," *Physiological measurement*, vol. 34, no. 7, p. R29, 2013.

[72] M. de Zambotti, F. C. Baker, and I. M. Colrain, "Validation of sleep-tracking technology compared with polysomnography in adolescents," *Sleep*, vol. 38, no. 9, pp. 1461–1468, 2015.

[73] S. Bhat, A. Ferraris, D. Gupta, M. Mozafarian, V. A. DeBari, N. Gushway-Henry, S. P. Gowda, P. G. Polos, M. Rubinstein, H. Seidu *et al.*, "Is there a clinical role for smartphone sleep apps? comparison of sleep cycle detection by a smartphone application to polysomnography," *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, vol. 11, no. 7, p. 709, 2015.

[74] D. J. McIver, J. B. Hawkins, R. Chunara, A. K. Chatterjee, A. Bhandari, T. P. Fitzgerald, S. H. Jain, and J. S. Brownstein, "Characterizing sleep issues using twitter," *Journal of medical Internet research*, vol. 17, no. 6, 2015.

[75] M. Almalki, K. Gray, and F. Martin-Sanchez, "Activity theory as a theoretical framework for health self-quantification: a systematic review of empirical studies," *Journal of medical Internet research*, vol. 18, no. 5, 2016.

[76] C. Paton, M. Margaret, L. Fernandez-Luque, and A. Y. Lau, "Self-tracking, social media and personal health records for patient empowered self-care," 2012.

[77] K. E. Heron and J. M. Smyth, "Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments," *British journal of health psychology*, vol. 15, no. 1, pp. 1–39, 2010.

[78] T.-C. Lu, C.-M. Fu, M. H.-M. Ma, C.-C. Fang, and A. M. Turner, "Healthcare applications of smart watches: a systematic review," *Applied clinical informatics*, vol. 7, no. 3, p. 850, 2016.

[79] H. E. Montgomery-Downs, S. P. Insana, and J. A. Bond, "Movement toward a novel activity monitoring device," *Sleep and Breathing*, vol. 16, no. 3, pp. 913–917, 2012.

[80] W. Knight, "The dark secret at the heart of al," *TECHNOLOGY REVIEW*, vol. 120, no. 3, pp. 54–61, 2017.

[81] A. Sathyanarayana, F. Ofli, L. Fernandez-Luque, J. Srivastava, A. Elmagarmid, T. Arora, and S. Taheri, "Robust automated human activity recognition and its application to sleep research," in *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on.* IEEE, 2016, pp. 495–502.

[82] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 219–238, 2016.

[83] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2017.

[84] I. Atmosukarto, N. Ahuja, and B. Ghanem, "Action recognition using discriminative structured trajectory groups," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on.* IEEE, 2015, pp. 899–906.

[85] S. Ma, J. Zhang, S. Sclaroff, N. Ikizler-Cinbis, and L. Sigal, "Space-time tree ensemble for action recognition and localization," *International Journal of Computer Vision*, pp. 1–19, 2017.

[86] Z. Zhou, F. Shi, and W. Wu, "Learning spatial and temporal extents of human actions for action detection," *IEEE Transactions on multimedia*, vol. 17, no. 4, pp. 512–525, 2015.

[87] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[88] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2593–2600.

[89] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *European Conference on Computer Vision*. Springer, 2014, pp. 581–595.

[90] H. Kuehne, J. Gall, and T. Serre, "An end-to-end generative framework for video segmentation and recognition," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–8.

[91] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[92] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[93] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4597–4605.

[94] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[95] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 11, pp. 1993–2008, 2013.

[96] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu, "Recognizing actions from still images," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on.* IEEE, 2008, pp. 1–4.

[97] C. Thurau and V. Hlavác, "Pose primitive based human action recognition in videos or still images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8.

[98] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg, "Coloring action recognition in still images," *International journal of computer vision*, vol. 105, no. 3, pp. 205–221, 2013.

[99] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.* IEEE, 2010, pp. 2030–2037.

[100] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 2011, pp. 3177–3184.

[101] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Computer Vision (ICCV), 2011 IEEE International Conference on.* IEEE, 2011, pp. 1331–1338.

[102] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for semantic description of humans in still images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 87–101, 2017.

[103] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343–3361, 2014.

[104] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[105] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[106] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*. IEEE, 2012, pp. 14–19.

[107] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014.

[108] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3d discriminative skeletal features for human action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 471–478.

[109] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[110] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[111] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." in *IJCAI*, vol. 13, 2013, pp. 2466–2472.

[112] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," *Computer Vision–ECCV 2006*, pp. 359–372, 2006.

[113] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 144–149.

[114] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Computer vision–ECCV 2012*. Springer, 2012, pp. 872–885.

[115] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2012, pp. 252–259.

[116] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[117] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, 2012.

[118] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[119] A. Calzada, J. Liu, C. Nugent, H. Wang, and L. Martinez, "Sensor-based activity recognition using extended belief rule-based inference methodology," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 2694–2697.

[120] D. Biswas, A. Cranny, N. Gupta, K. Maharatna, J. Achner, J. Klemke, M. Jöbges, and S. Ortmann, "Recognizing upper limb movements with wrist worn inertial sensors using k-means clustering classification," *Human movement science*, vol. 40, pp. 59–76, 2015.

[121] A. M. Tripathi, D. Baruah, and R. D. Baruah, "Acoustic sensor based activity recognition using ensemble of one-class classifiers," in *Evolving and Adaptive Intelligent Systems (EAIS), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–7.

[122] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition," *Applied Soft Computing*, vol. 37, pp. 1018–1022, 2015.

[123] R. Akhavian and A. Behzadan, "Wearable sensor-based activity recognition for data-driven simulation of construction workers' activities," in *Winter Simulation Conference (WSC), 2015*. IEEE, 2015, pp. 3333–3344.

[124] L. Liu, Y. Peng, M. Liu, and Z. Huang, "Sensor-based human activity recognition system with a multilayered model using time series shapelets," *Knowledge-Based Systems*, vol. 90, pp. 138–152, 2015.

[125] G. Okeyo, L. Chen, H. Wang, and R. Sterritt, "Dynamic sensor data segmentation for real-time knowledge-driven activity recognition," *Pervasive and Mobile Computing*, vol. 10, pp. 155–172, 2014.

[126] B. H. Dobkin, "Wearable motion sensors to continuously measure real-world physical activities," *Current opinion in neurology*, vol. 26, no. 6, p. 602, 2013.

[127] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[128] J. Bort-Roig, N. D. Gilson, A. Puig-Ribera, R. S. Contreras, and S. G. Trost, "Measuring and influencing physical activity with smartphone technology: a systematic review," *Sports Medicine*, vol. 44, no. 5, pp. 671–686, 2014.

[129] C. S. Liew, T. Y. Wah, J. Shuja, B. Daghighi *et al.*, "Mining personal data using smartphones and wearable devices: A survey," *Sensors*, vol. 15, no. 2, pp. 4430–4469, 2015.

[130] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE sensors journal*, vol. 15, no. 3, pp. 1321–1330, 2015.

[131] L. M. Zhou, C. Gurrin, and Z. Qiu, "Zhiwo: Activity tagging and recognition system for personal lifelogs," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 321–322.

[132] C. Meurisch, B. Schmidt, M. Scholz, I. Schweizer, and M. Mühlhäuser, "Labels: Quantified self app for human activity sensing," in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 1413–1422.

[133] J. Hamm, B. Stone, M. Belkin, and S. Dennis, "Automatic annotation of daily activity from smartphone-based multisensory streams," in *International Conference on Mobile Computing, Applications, and Services*. Springer, 2012, pp. 328–342.

[134] C. Dobbins and R. Rawassizadeh, "Clustering of physical activities for quantified self and mhealth applications," in *Computer and Information Technology; Ubiquitous Computing and*

Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on. IEEE, 2015, pp. 1423–1428.

[135] M. Uddin, A. Salem, I. Nam, and T. Nadeem, "Wearable sensing framework for human activity monitoring," in *Proceedings of the 2015 workshop on Wearable Systems and Applications.* ACM, 2015, pp. 21–26.

[136] O. Banos, J. Bang, T. Hur, M. H. Siddiqi, H.-T. Thien, L.-B. Vui, W. A. Khan, T. Ali, C. Villalonga, and S. Lee, "Mining human behavior for health promotion," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE.* IEEE, 2015, pp. 5062–5065.

[137] N. James and D. Matteson, "ecp: An r package for nonparametric multiple change point analysis of multivariate data," *Journal of Statistical Software*, vol. 62, no. 1, pp. 1–25, 2015.

[138] R. P. Troiano, D. Berrigan, K. W. Dodd, L. C. Masse, T. Tilert, M. McDowell *et al.*, "Physical activity in the united states measured by accelerometer," *Medicine and science in sports and exercise*, vol. 40, no. 1, pp. 181–188, 2008.

[139] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[140] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 881–892, 2002.