# DISEASE STATE PREDICTION USING MULTISCALE DYNAMICS

A Dissertation
Presented to
The Academic Faculty

By

Ayse Selin Cakmak

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering
College of Engineering

Georgia Institute of Technology

August 2021

# DISEASE STATE PREDICTION USING MULTISCALE DYNAMICS

Thesis committee:

Dr. Gari D. Clifford
The Wallace H. Coulter Department of Biomedical Engineering
*Georgia Institute of Technology*

Dr. Christopher J. Rozell
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. Omer T. Inan
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. David V. Anderson
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. Eva Dyer
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. Amit J. Shah
Department of Epidemiology, Rollins School of Public Health
*Emory University*

Date approved: 06/08/2021

If you were born with the weakness to fall, you were born with

the strength to rise.

*Rupi Kaur*

To my parents Serap and Erdal Cakmak; my grandparents Necla and Atalay Kaya, Gulsevin and Ahmet Cakmak

# ACKNOWLEDGMENTS

Many people describe Ph.D. as a journey, and I can understand them better when it is finally time for me to write this section of the dissertation. As with any great journey, there are ups and downs, getting lost or taking breaks. When the path is this difficult, reaching the other end is impossible without having companions and wise mentors. I have been beyond lucky to have many great friends, mentors, and family who have helped me complete my thesis.

First and foremost, I would like to thank my Ph.D. advisor, Professor Gari Clifford. He is one of the pioneers in our field; I have been beyond lucky to be mentored by this world-renowned researcher. Even though he has the busiest schedule as the department chair and jointly appointed professor, he has always been accessible and approachable. During our meetings, he had motivated me and supported me when I wanted to pursue my ideas. His knowledge and ideas have inspired me countless times over the years. He was one of the reasons that our lab group is so close together. I will always cherish the memories of our department parties, Euro Cups at the department, and lab outings. I want to thank him so much for taking a chance on me and accepting me as his student in 2017 winter. I will keep on following his path and his example.

I would like to thank my co-advisor, Prof. Rozell. He is very skillful at giving feedback on research topics. I had the opportunity to ask for his feedback as a co-advised student. I would also like to thank my thesis committee Dr. Amit Shah, Dr. Omer Inan, Dr. Eva Dyer, and Dr. David Anderson, for taking the time to serve on my dissertation committee. Their valuable comments and suggestions have improved my work. I have done two great internships during my Ph.D years and had the chance to work with amazing researchers in the m-health field. I would like to thank my managers and mentors; Dr. Nina Thigpen, Pramod

Gupta, Obi Felten, Dr. Chris Paetsch, and Tegan Ayers. I learnt a lot from them during 2019 and 2020 summers.

I was so lucky to have such great labmates who became family over the years. Dr. Giulia Da Poian, Dr. Justus Schwabedal, Dr. Erick Perez Alday, Dr. Ali Bahrami Rad were postdocs during my time at the Clifford lab. They gave me much helpful feedback and lots of support whenever I needed it. Dr. Erik Reinertsen was my first labmate and mentor; he taught me a lot in his last year and made sure I had a great foundation to start on. I have spent so many brainstorming sessions, outings, coffee chats with my labmates Pradyumna Suresha, Nick Shu, Samaneh Nasiri, Chaitra Hedge, Camilo Valderrama, Zifan Jiang, Sam Waters, and Nasim Katebi. I will never forget the times I spend with you, our lunches at Emory, and our international restaurant trips.

I have met amazing friends outside the lab too. Dr. Bige Deniz Unluturk was my dearest roommate, one of my best friends, and sometimes mentor in life and research. During my Ph.D, it was such a pleasure to meet and become friends with Emre Yilmaz, Dr.s Beren and Emre Gursoy, Dr. Nil Gurel, Guliz and Goktug Ozmen, Mine and Mustafa Sak, Dr. Gokcin Cinar, Efe Yarbasi, Deniz Uysal, and Ezgi Balkas. You have made Georgia Tech so fun, and it was such a pleasure to cross paths with you in the US.

I want to thank my love, the most inspiring researcher, the kindest soul I met during this journey. No words can express how thankful I am to Dr. Dogancan Temel. We are from the same high school and studied in the same classroom, but it was in our stars to meet in Atlanta, in another part of the world. When I was down, he lifted me and made me see everything in a different light. When I was lost, he guided me and gave me the best advice. Being my senior and from the same department, I came to his door with many research questions. Meeting him was one of the best things that happened in my life. He made my dreams come true. To many more journeys and adventures together!

I want to thank my parents and grandparents, to whom I have dedicated this thesis. My mom and dad, Serap and Erdal Cakmak have been my most significant source of support always. I have talked to them countless times over the Ph.D. journey, telling stories of minor and major challenges, and they never stopped believing in me or supporting me. They have always sent me to the best schools, provided me with the best opportunities, and stood behind me always. My grandparents, Necla Kaya, Atalay Kaya, Gulsevin and Ahmet Cakmak have supported my dreams unconditionally, even though it sometimes meant I would be spending many years away from them. Whenever I called them, they gave me words of encouragement and love. I would not have been writing these words if this was not for you. Unfortunately, my grandfather Ahmet Cakmak passed away one week before my defense. I am sure he watched me from above during this final part of my Ph.D. His memory will always live with me in my heart; may he rest in peace.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ACRONYMS

**6MWD** 6-minute walk distance

**6MWT** 6-minute walk test

**AHI** Apnea–hypopnea index

**AMoSS** Automated Monitoring of Symptom Severity

**AUC** Area Under The Curve

**AUCPr** Area under the precision-recall curve

**AURORA** Advancing Understanding of RecOvery afteR traumA

**BBlocks** Bayesian Blocks

**BiS** Binary Segmentation

**BOCPD** Bayesian Online Change Point Detection

**CPD** Change Point Decoder

**DIMS** Disorders of Initiating and Maintaining Sleep

**DMD** Duchenne muscular dystrophy

**DSM-5** Diagnostic and Statistical Manual of Mental Disorders

**ED** emergency department

**GLM** Generalized Linear Model

**HF** Heart Failure

**HPA** Hypothalamic-Pituitary-Adrenal

**HRV** Heart Rate Variability

**IS** Interdaily Stability

**IV** Intradaily Variability

**K-NN** K-Nearest Neighbors

**KCCQ** Kansas City Cardiomyopathy Questionnaire

**L5** Least Active 5 Hours

**LOOCV** leave-one-out cross validation

**M10** Most Active 10 Hours

**MAE** Mean Absolute Error

**mBOCPD** Modified Bayesian Online Change Point Detection

**MCEPS** Michigan Critical Events Perception Scale

**MLP** Multilayer Perceptron

**NN** Normal-To-Normal

**NYHA** New York Heart Association

**OA** Oakley

**PCL-5** PTSD Checklist for DSM-5

**PDI** Peritraumatic Distress Inventory

**PELT** Pruned Exact Linear Time

**PLMI** periodic limb movement index

**PPG** Photoplethysmogram

**PPV** Positive predictive value

**PROMIS** Patient-Reported Outcomes Measurement Information System

**PSG** Polysomnography

**PSQIA** Pittsburgh Sleep Quality Index Addendum

**PTSD** Post-traumatic Stress Disorder

**RMDM** Recursive Mean Difference Maximization

**ROC** Receiver Operating Characteristic

**SCN** suprachiasmatic nucleus

**SE** Sleep Efficiency

**SHAP** SHapley Additive exPlanation

**SVM** Support Vector Machine

**TPR** True positive rate

**TV** Total Variation

**VAE** variational autoencoder

**WASO** Wake After Sleep Onset

# SUMMARY

Human physiology shows a wide range of cyclical physiological changes coordinated by endogenous biological clocks. These clocks work in different time scales. Circadian rhythms follow the earth's day-night cycle and have a period of 24 hours, while the ultradian rhythms have shorter periods. Multiscale changes and variations in our physiology and behavior ensure optimal health and adaptation to our environment.

Disruptions in physiological rhythms have been shown to reflect illness, or indeed, can lead to or exacerbate underlying conditions. For example, it has been shown that the likelihood of suffering cardiovascular complications such as sudden cardiac death or arrhythmia is associated with the circadian rhythm [1]. This may reveal a circadian mechanism of action. It is well-known that poor sleep can lead to a variety of mental health issues [2]. Moreover, both physical and mental health disorders (such as cardiovascular disease and depression) can lead to circadian disturbances, including difficulties falling asleep, nocturnal panic attacks, and insomnia [3, 4]. This can create a reinforcement feedback loop to exacerbate both mental and physical health issues.

Treatment of many disorders requires clinic visits, with a strong reliance on patient self-reports for filling in the gaps between these in-person assessments. This lack of monitoring leads to both biases in the data (self-reporting/recollection is well-known to be problematic [5]) and a lack of provision for responding to rapid changes in health (such as cardiovascular decompensation or suicide risk). Objective tools for monitoring patients in between clinical visits are largely absent. In recent years, advances in (and a push in commercialization) of wearable technology have enabled almost real-time monitoring of changes in physiology. This has the potential to transform the landscape for monitoring diseases. However, despite these advances, there has been a general lack of rigorous studies ‘in the wild’ and a lack of focus on novel metrics derived from the raw data.

The research described in this thesis aims to address the gap in the area of passive and continuous monitoring using wearables in naturalistic settings. Towards this goal, the focus of this research is on building signal processing and machine learning frameworks to quantify non-stationarities in physiological signals and biological rhythm disruptions. In the first part of this dissertation, a wearable-based sleep detection approach is developed. The proposed approach uses the variations observed in the motion and heart rate data and detects patterns in these change events associated with sleep-wake transitions. By combining different modalities, this approach achieved higher wake detection accuracy compared to a solely actigraphy-based method on a clinical cohort.

The second part of this dissertation focuses on features quantifying biological rhythms to separate healthy controls and participants with health disorders and validates the developed techniques in two different applications. The first work presents the use of accelerometer and Photoplethysmogram (PPG) signals to derive health outcomes post-trauma, based on analyzing circadian and ultradian rhythm features with machine learning algorithms. Then, deep learning algorithms are employed to extract features and representations for the same classification task. The second work presents the use of both passive (motion, location, social contact) and active (clinically-validated survey) data collected by a smartphone app for monitoring HF patients non-invasively. The results show that these data modalities could provide a complementary continuous monitoring approach.

Once verified by deploying on devices in real-life settings, the tools in this dissertation can potentially assist caregivers in monitoring patients and inform more timely and appropriate clinical decisions and interventions.

# CHAPTER 1
# INTRODUCTION

Wearables could improve upon conventional long-term patient monitoring tools because they can collect diverse and objective data types, provide insights into patients' behavior in naturalistic settings, and capture biological rhythms over multiple time scales ranging from seconds to weeks. Treatment approaches for many disorders are in-person clinic visits and monitoring patients with self-reported questionnaires in between the assessments. However, varying symptoms might not be captured in one hospital visit's data [6], or might not reflect the patient's behavior in the home environment [7]. While administering questionnaires in the gaps between the clinic visits aims to solve these problems, these self-reporting tools are subjective and prone to recall bias [5]. Furthermore, some disorders may require dynamic monitoring of changes to provide timely care and support. Developing novel wearable-based monitoring methods would enable objective longitudinal assessments over multiscale physiological dynamics.

Biological rhythms occur on different time scales and have diverse cycle lengths. The biological clock that follows the earth's day-night cycle and has 24 hours period is called the circadian rhythm [8]. The suprachiasmatic nucleus (SCN) of the hypothalamus is the master clock that controls the circadian oscillators [3] and regulates the synthesis of melatonin [9]. Therefore, the circadian master clock is involved in orchestrating the timing and the structure of the sleep/wake cycle. Ultradian rhythms result from various biological processes and have periods less than 24 hours [10]. Some examples of ultradian rhythms include sleep states, electrical activity changes in the brain, respiration, and circulation.

While the biological clocks are essential mechanisms for our survival and adaptation to our environment, underlying adverse conditions could disrupt these mechanisms. In the previous studies, it has been shown that physical and mental health disorders could lead to circadian rhythm disturbances, including difficulties falling or staying asleep, nocturnal panic attacks, or more severe symptoms at particular times of the day [3]. While these circadian rhythm disturbances are considered as symptoms of disorders, there is a bidirectional relationship, and heightened circadian disturbances could exacerbate the symptoms of mental and physical health issues [2]. Furthermore, it has been shown that cardiovascular events such as cardiac death or arrhythmia demonstrate circadian rhythms [1], and Heart Rate Variability (HRV) metrics follow circadian and ultradian cycles [11]. Participants with Post-traumatic Stress Disorder (PTSD) show alterations in the Hypothalamic-Pituitary-Adrenal (HPA) axis which regulates ultradian rhythms, and therefore have an impact the cardiovascular system [12]. One of the primary symptoms of the disorder is hyperarousal events, resulting in sudden heart rate elevations. Consequently, studying the biological rhythm dysregulation over days as well as smaller timescales could provide novel insights into both cardiovascular and mental health disorders and may widen our comprehension of their impact on physical health.

This dissertation describes the use of various data modalities from wearable devices to build low-cost methods for passive monitoring. These methods can complement conventional approaches (such as self-report questionnaires) and improve the disease state estimation performance. In this work, wearable data is analyzed in multiple timescales. On the ultradian scale, a novel sleep detection technique was developed, and HRV metrics were derived. The circadian rhythm was investigated using standard features and unsupervised machine learning approaches to generate novel features. Machine learning models were built to estimate or predict disease states from these features. Figure 1.1 illustrates some examples of ultradian and circadian cycles as well as the disease states. The upper plot shows the sleep/wake cycle, an ultradian rhythm occurring on an hourly time scale. In the middle plot, daily changes in activity are shown to illustrate the circadian rhythm. The bottom plot shows the disease states, which tend to move from an acute phase to a chronic phase.

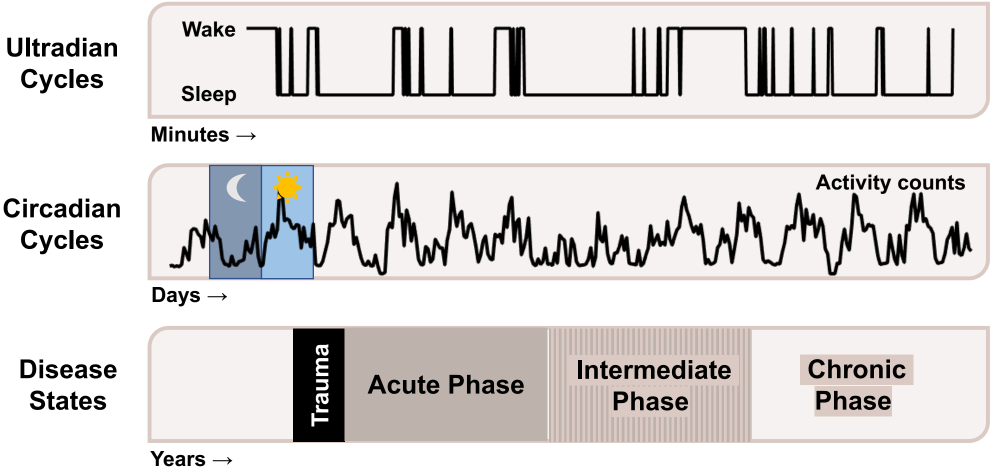


Figure 1.1: Biological rhythms and physiological states on three temporal scales.

## 1.1 Major contributions

The major contributions of this work can be summarized as follows:

- Designed and implemented a novel sleep/wake state detection method from movement and physiological signals collected from wearable devices.

- Designed and implemented a machine learning-based algorithm for mapping motion and heart rate variability features that capture circadian and ultradian variability from longitudinal wearable data to estimate post-trauma outcomes.

- Developed a novel method for unsupervised feature extraction from 2-dimensional representations of movement data using deep learning.

- Designed and implemented a method for using passively collected smartphone data to predict heart failure decompensation events and developed a novel late-fusion approach to fusing multimodal data.

The above contributions have been published in the following:

**Journal articles:**

- Siegel, B.I., **Cakmak, A.S.**, Reinertsen, E., Benoit, M., Figueroa, J., Clifford, G.D. and Phan, H.C., 2020. Use of a wearable device to assess sleep and motor function in Duchenne muscular dystrophy. Muscle & Nerve, 61(2), pp.198-204.

- **Cakmak, A.S.**, Da Poian, G., Willats, A., Haffar, A., Abdulbaki, R., Ko, Y.A., Shah, A.J., Vaccarino, V., Bliwise, D.L., Rozell, C. and Clifford, G.D., 2020. An unbiased, efficient sleep-wake detection algorithm for a population with sleep disorders: change point decoder. Sleep, 43(8).

- **Cakmak, A.S.**, *et al.* 2021. Classification and prediction of post-trauma outcomes related to PTSD using circadian rhythm changes measured via wrist-worn research watch in a large longitudinal cohort. IEEE Journal of Biomedical and Health Informatics [in press].

**Conference articles and abstracts:**

- **Cakmak, A.S.**, Reinertsen, E., Taylor, H.A., Shah, A.J. and Clifford, G.D., 2018, December. Personalized heart failure severity estimates using passive smartphone data. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 1569-1574). IEEE.

- **Cakmak, A.S.**, Lanier, H.J., Reinertsen, E., Harzand, A., Zafari, A.M., Hammoud, M.A., Alrohaibani, A., Wakwe, C., Appeadu, M., Clifford, G.D. and

Shah, A.J., 2019. Passive Smartphone Actigraphy Data Predicts Heart Failure Decompensation. Circulation, 140(Suppl—1), pp.A15444-A15444.

- **Cakmak, A.S.**, Thigpen, N., Honke, G., Alday, E.P., Rad, A.B., Adaimi, R., Chang, C.J., Li, Q., Gupta, P., Neylan, T., and McLean, S.A., 2020. Using Convolutional Variational Autoencoders to Predict Post-Trauma Health Outcomes from Actigraphy Data. Presented at NeurIPS (Machine Learning for Mobile Health), Dec 12, 2020 (online), also available as an arXiv preprint arXiv:2011.07406. [Selected as spotlight presentation.]

The following article is currently in submission:

- **Cakmak, A.S.**, Shah, A.J. and Clifford, G.D. Dynamics of interpersonal social interactions and motion passively captured from smartphones predict decompensation in heart failure.

### 1.2 Scope and organization of the dissertation

This work is organized as follows: Chapter 2 presents background on sleep detection on wearables. Following this, PTSD, heart failure, and Duchenne muscular dystrophy (DMD) disorders are described in detail. The chapter concludes with detailed descriptions of the datasets used in this dissertation.

Chapter 3 focuses on variations on shorter time scales and ultradian rhythms. The chapter summarizes change point detection techniques and assesses their performance on artificial heart rate data. Then, a novel sleep detection technique that utilizes change points observed in physiological and movement signals from wearables is introduced. The technique's performance was measured on a dataset with simultaneous wearable, and Polysomnography (PSG) recordings from a sleep clinic and the expert annotations from the sleep clinic were used as ground truth. The ability of the technique to accurately estimate sleep study parameters such as sleep efficiency was also assessed and compared to a movement data-based approach. Finally, performance on a second dataset with a second research watch was presented to demonstrate the generalizability of the proposed technique to different devices.

Chapter 4 discusses the use of HRV and movement metrics from the research watch for monitoring patients post-trauma passively. Previously validated and novel features to quantify the ultradian and circadian variability are derived from longitudinal watch data, and machine learning techniques are used to map these to clinical outcomes. The second project discusses using an unsupervised deep learning method to derive features from movement data and compare performance with a fully supervised approach. Furthermore, sleep detection described in Chapter 3 is applied to the data collected in daily living conditions to capture the sleep disturbances post-trauma.

In Chapter 5, the analysis for intermediate and chronic disorders is presented. In the first project, a smartphone-based framework is used to collect data for heart failure patients. The HF decompensation event prediction ability of various single-modality models are compared. It is shown that the late fusion approach improves the performance of the models, and a time-to-event analysis is presented. In the following sub-sections, circadian rhythm metrics are derived from movement data of intermediate and chronic state PTSD, and DMD patients and relation to clinical outcomes are assessed. Finally, Chapter 6 discusses concluding remarks and future directions.

# CHAPTER 2
# BACKGROUND

Smartphones and wearables become a part of our daily lives and give us an amazing opportunity to collect physiological data, in every part of the life and during many human activities. In this part of the thesis, wearable-based methods for sleep detection and disease state estimation (for post-traumatic stress disorder and heart failure) are reviewed.

## 2.1 Sleep states and wearable-based sleep/wake detection

Several sleep/wake classification algorithms for wearables have been suggested over the last decades, and they are typically based solely on actigraphy derived from accelerometer [13, 14, 15, 16]. Several findings suggest that only using movement signals leads to the main limitation of current algorithms: the incorrect classification and overestimation of low activity tasks as such sleep [17, 18, 19]. Indeed, low activity (quiescent) segments are not unique to sleep but are common to other activities such as reading or watching television. Another limitation results from the adoption of imprecise evaluation metrics used in assessing the performance of these devices. Since the percentage of sleep is typically higher compared to wake overnight, total accuracy may not be a reliable metric to evaluate performance. Sleep/wake detection may be considered as a "rare class problem" and may be amenable to alternative model evaluation metrics which better reflect this issue.

The first approaches in the field for state determination were based on calculating a weighted sum over the actigraphy epochs around the current epoch and scaling the summation to distinguish sleep from wakefulness [13, 14]. Oakley presented a similar approach in which the current epoch, epochs in the 2 minutes before and the 2 minutes after the current epoch are scaled with predetermined coefficients and summed [20]. If the summation is higher than the threshold, the region was labeled as wake. The Oakley algorithm is utilized in commercially available devices with different threshold selections (e.g., Actiwatch 2, Philips Respironics; Bend, Oregon). These actigraphic methods rely solely on the amplitude of actigraphic signals, which makes them low cost and easy to implement. However, these methods may overestimate sleep, particularly for patients with disordered sleep [21, 22]. It has been long been known that heart rate reflects transitions from sleep to wake and from wake to sleep [23, 24, 25]. Recent studies in the field leverage a combination of photoplethysmography (PPG) and accelerometer signals for sleep/wake detection [26, 27, 28]. However, these approaches have not been tested on clinical populations and still show low sensitivity in detecting wake epochs.

## 2.2 Post-traumatic stress disorder

PTSD is a psychiatric condition that can develop after exposure to threatening or horrifying events. Significant symptoms consistent with the eventual development of PTSD may manifest within days, weeks or months, and more rarely, a year or two after the traumatic event [29]. Symptoms may include persistent intrusive memories of trauma, sleep disturbances, avoidance of stimuli related to the trauma, hyperarousal, and negative changes in mood and cognition. PTSD can result from events such as violent personal assaults, natural or human-caused disasters, motor vehicle collisions, combat, and other forms of violence [30]. It has been shown that patients with PTSD experience sleep disturbance, particularly in terms of nightmares and panicked awakenings from sleep [4]. In addition, various studies suggest a significant comorbidity of pain with PTSD [31]. Many models have been developed to explain this co-occurrence of pain and PTSD, including the mutual maintenance model [32]. According to this model, pain acts as a reminder of the traumatic event and maintains PTSD symptoms. Then, these symptoms reduce the ability to cope with pain effectively. Although approximately 90% of all U.S. adults report exposure to at least one traumatic event in their lifetime, most do not develop PTSD [33]. It has been shown in previous studies that the majority of individuals experience PTSD onset within the first three months after trauma, while "delayed expression" PTSD (after six months) was observed on average for 15.3% of the cases [34].

PTSD prediction using standard survey data remains a challenge, since potential risk factors (such as age, gender, previous trauma) did not show a strong association with PTSD [35, 36, 37]. In a previous study, Schultebraucks *et al.* [38] combined biomarker data with clinical assessments from the emergency department (ED) to build a cross-validated prediction algorithm. By fusing these two modalities, the model's Area Under The Curve (AUC) for classifying participants with non-remitting PTSD symptoms from participants with resilient trajectories was 0.83 on a validation dataset. They also tested the use of electronic medical records alone and achieved an AUC of 0.72, which outperformed the baseline classifier (AUC=0.62). In another work, video and audio-based features were used with a deep learning classifier and achieved an AUC of 0.90 for predicting PTSD one month after ED enrollment [39].

The exponential increase in consumer wearables, and in wearable technology generally, has created an exciting opportunity to predict adverse mental health outcomes using wrist-wearable data [40, 41]. Two key outputs of wrist-wearable data are HRV and actigraphic data. Individual differences in a various time- and frequency-domain HRV measures have been found to predict a range of mental and physical health outcomes, including depression, anxiety, and poorer cardiovascular health [42, 43]. On the other hand, individuals with established PTSD have been shown to have HRV profiles consistent with increased sympathetic nervous system activity during sleep [44, 45]. In a previous pilot study, by using a dataset of 23 subjects with current PTSD and 25 control subjects, the authors found that HRV features derived from time periods with the lowest heart rate in 24-hour periods classify PTSD with an AUC of 0.86 [46]. McDonald *et al.* used heart rate data from 100 participants to detect the onset of PTSD triggers [47]. By combining the heart rate features with and Support Vector Machine (SVM), authors achieved an AUC of 0.67 and found that their algorithms associated the increase of heart rate with onset of PTSD trigger.

Motion data provides a cost-effective solution for monitoring participants over extended periods of time. It can be used to estimate sleep disturbance using derived sleep/wake estimates and the rest/activity patterns [48]. Many studies utilized actigraphy as an objective tool to characterize disturbances in sleep and circadian rhythm in PTSD [49, 5]. However, analyses were confined to identifying statistically significant differences in populations and cross-validated classification analysis was not performed. In another study, Tsanas *et al.* developed a sleep period estimation algorithm using accelerometer and ambient light data from wearables [50]. Authors find that their proposed method can estimate the sleep onset and sleep offset data captured by self-reported sleep diaries on a cohort including 42 PTSD participants. In addition, they inspect the group differences and found that the activity during sleep and Intradaily Variability (IV) features to be statistically significant across groups. However, to author's knowledge, no other study has combined heart rate and motion data to build an objective multivariate classifier over extensive periods of time. This dissertation combines both modalities to analyze multiscale biological rhythms and builds objective passive monitoring tools.

## 2.3 Heart failure

The American Heart Association estimates that between 2013 and 2016, approximately 6.2 million Americans had HF [51]. In 2012, the economic burden of HF was estimated at $30.7 billion. Projections suggest a 127% increase in cost by 2030. Overall, cardiovascular diseases account for the highest expenditures amongst all non-communicable diseases in the US [52].

HF decompensation, associated with hypervolemia (volume overload), is defined as a clinical syndrome in which a functional change in the heart leads to new or increasing symptoms, including fatigue, dyspnea, and edema, and requires hospitalization [53]. Treatment includes diuretics and vasodilators intended to improve volume status and cardiac function. Unfortunately, even following successful treatment and return to the euvolemic (normal volume status) state, decompensation episodes can continue to occur with increasing frequency [53, 54]. Patil *et al.* reported that about 20% of the patient cohort were readmitted within 30 days of initial hospitalization due to HF, with a median readmission time of 12 days [55]. Furthermore, patients with a lower income had a higher readmission rate, indicating that socio-economical factors could also contribute to the disease's progression. If low-cost monitoring methods identify decompensation episodes developing outside the clinic, medical interventions could be administered proactively to prevent hospitalization or other adverse outcomes.

Various studies investigated techniques for monitoring HF patients non-intrusively. Packer *et al.* [56] showed that using a combination of clinical variables and impedance cardiography features could be a predictor of a decompensation event in the next 14 days. Previous studies have also investigated the use of wearable devices adhered to the chest. In the 'Multisensor Monitoring in Congestive Heart Failure' study, the authors propose an algorithm that uses physiological signals, and they report a sensitivity of 63%, and specificity of 92% [57]. However, the authors provide few details and claim it is 'proprietary'. Inan *et al.* recorded seismocardiogram signal with a non-invasive wearable patch before and after a 6-minute walk test to analyze the cardiac response to exercise [58]. The authors used graph similarity scores between the rest and recovery phases and found a significant difference between compensated and decompensated groups. In another example, similarity-based modeling was used with physiological signals from a patch on the chest to detect changes from the baseline. This algorithm had a sensitivity of 88% and specificity of 85% [59]. Using ballistocardiogram data recorded at home was also investigated [60], and authors demonstrated that collecting high-quality ballistocardiogram data at home is feasible, and an AUC of 0.78 could be achieved for classifying clinical status. Other non-invasive approaches include patient-reported outcomes, which could be collected using clinically validated questionnaires such as KCCQ. The KCCQ assesses the quality of life, predict readmissions and mortality in HF patients [61]. In a previous study, Flynn *et al.* reported that KCCQ has modest correlations with exercise capacity measured by the 6-minute walk test in a population with HF [62].

With the advancement of technology, smartphones have become a ubiquitous part of our daily life. For long-term monitoring, using a smartphone could be advantageous to a solution requiring an additional device by reducing the disruption to patients' normal daily routine. Our research team and collaborators have previously developed the Automated Monitoring of Symptom Severity (AMoSS) app, which is a custom and scalable smartphone-based framework for remote monitoring [63]. Subsequently, the current authors used the passive data from the first ten participants of this study to estimate the KCCQ surveys collected through the app [64]. The model estimated the KCCQ score with a mean absolute error of 5.7%, providing an entirely passive method of monitoring HF related quality of life. (The method was passive in the sense that it does not require any active participation by either the patient or clinical staff beyond the everyday use of a mobile phone to monitor activity and behavioral patterns in the background using software.) Then, in subsequent work, motion data was used to classify decompensation or compensation events [65]. By using a hold-out test randomly sampled from 30% of the events ($N_{test} = 32$), the AUC of the classifier was found to be 0.76. Heart failure decompensation events were also predicted from features derived from passive and active data collected by the smartphone-based framework [66]. Features were extracted from multiple modalities including motion, social contact, location, and clinical survey data (KCCQ). Algorithms based on using a single modality and two different sensor fusion approaches were developed. An analysis of the feature importance in the model is also presented. Finally, a novel late-fusion model that combines the KCCQ, motion, and social contact data is proposed.

## 2.4 Increased risk of cardiovascular disease due to post-traumatic stress disorder

PTSD has been associated with cardiovascular diseases and previous studies have shown that PTSD might have harmful effects on cardiovascular health due to following reasons:

- **Physical changes:** One of the main symptoms of PTSD include hypervigilance, and the reminders of the traumatic event results in higher sympathetic nervous system activity [12]. The activity of is sympathoadrenal axis is also increased, and this leads to higher levels of catecholamine, which could effect the heart [67].
- **Lifestyle factors:** PTSD symptoms could lead to unhealthy behaviours in-

cluding smoking, alcoholism, or physical inactivity [68]. In addition, obesity might also have an association with PTSD and patients with PTSD report higher Body Mass Index.

Vaccarino *et al.* used a dataset of male twins from the Vietnam Era Twin Registry to analyze the association of PTSD with coronary heart disease [12]. Authors found that the incidence of heart disease is double in twins with PTSD. On another study, Kang *et al.* studied World War II prisoners of war, and found that cardiovascular disease risk is significantly higher if participant has PTSD [69].

## 2.5 Duchenne muscular dystrophy

DMD is a genetic disorder that results in progressive loss of functional muscle mass [70]. This progressive loss of muscle stats early in life and could result in the loss of life in adolescent years due to compromise of respiratory musculature. Symptoms of DMD include gait abnormalities, difficulty rising up from the ground, frequent falls, and sleep disorders [71, 72]. Previous studies have investigated utility of actigraphy data to estimate outcomes related to DMD. Davidson *et al.* used StepWatch activity monitor (Orthocare Innovations, Washington) to collect data from participants with DMD (N=16) and healthy controls (N=13) [73]. By using parameters derived from the watch, authors found participants with DMD had less step count and were inactive for longer periods. In another work, 22 participants with DMD wore an actigraph during their daily life [74]. The authors found the area under the curve for each 1-minute epoch of actigraphy data could be a useful feature to estimate the muscle strength. These works indicate that wearable-based monitoring could be useful for DMD, but metrics related to circadian rhythm have not been tested in this population yet.

## 2.6 Datasets used in this thesis

### 2.6.1 Emory post-traumatic stress disorder dataset

The dataset in Emory PTSD study includes a subgroup of participants (n = 102, men, mean age = 68.56, SD = 1.93) from the Emory Twin Study Follow-up recruited from the Vietnam Era Twin Registry [75]. Written informed consent was obtained from all participants, and the Emory University Institutional Review Board approved this research (IRB #00081004). All PSG data were collected from data acquisition systems (Natus, Remlogic) set up in two bedrooms in the Emory Sleep Center. During PSG, subjects wore a commercially available wrist-worn research watch (Empatica E4, Empatica; Cambridge, MA). The wrist-worn device recorded PPG and 3-axis accelerometer signals with sampling rates 64 Hz and 32 Hz respectively. After the participants were discharged from clinic, they start wearing Actiwatch research watch (Philips Respironics, Bend, Oregon). The device collected activity data continuously everyday, up to two weeks and aggregated into 30-second epochs.

### 2.6.2 Advancing Understanding of RecOvery afteR traumA post-traumatic stress disorder dataset

The Advancing Understanding of RecOvery afteR traumA (AURORA) dataset, used in this thesis, consisted of individuals who present to participating emergency departments within 72 hours of a traumatic event [76]. Traumatic events that qualified automatically for study enrollment were motor vehicle collision, physical assault, sexual assault, fall $>$ 10 feet, or mass casualty incidents. The patients ranged in age from 18 to 75 years. Although the AURORA study's aim is to collect data from 5000 individuals, the data is being analyzed in a series of tranches (or 'freezes') to report results to the scientific community. This approach also allows future data to act as a truly independent test set. For the current study, we present the analysis of the first set of participants (N=1618) enrolled between July 31, 2017, and July 31, 2019. There were 2312 subjects enrolled until July 31, 2019. Participants who were deceased, those who dropped, who were pregnant or incarcerated, or anyone for whom the medical data extraction form was not available were not included in the released analyzable cohort, making the final dataset size 1618 participants. These 1618 participants are referred to as 'Freeze 2' dataset. Demographics (age, sex, BMI, and employment status) of the participants are shown in Table 2.1.

Table 2.1: Freeze 2 dataset participant demographics. p values calculated using Wilcoxon rank sum test (age, BMI) or Fisher exact test (sex, employment) between PCL-5 $\geq 31$ and PCL-5 $< 31$ participants. Age and BMI are shown as Mean (SD).

| | | Total | Week-8 | Week-8 | p |
|---|---|---|---|---|---|
| | | | PCL-5 $\geq 31$ | PCL-5 $< 31$ | val. |
| Sex | M | 581 | 156 | 194 | 0.57 |
| | F | 1037 | 409 | 471 | |
| Age | | 35 (13) | 36 (12) | 35 (13) | 0.31 |
| BMI | | 30.4 (8.7) | 30.7 (9.2) | 30.3 (8.4) | 0.53 |
| Emp. Status | Employed | 1064 | 374 | 545 | 0.05 |
| | Other | 554 | 191 | 220 | |

The AURORA study protocol was ethically approved by the central Institutional Review Board (IRB #17-0703) at the University of North Carolina Chapel Hill. Participants were asked to wear a research watch (Verily Life Sciences; San Francisco, CA) at least 21 hours a day for the eight-week period and at subsequent times that vary by the study participant. This research watch collected accelerometry and the PPG data at 30 Hz sampling frequency.

### 2.6.3 Automated Monitoring of Symptom Severity heart failure dataset

AMoSS app, designed and implemented by our research team and collaborators, passively collects location, activity, clinical surveys and contact activity/diversity data via de-identified lists of word type, as well as recipients and senders of text messages and phone calls, including length/duration and time of day [77]. In order to protect subject's privacy, all data are de-identified at source, using hashed identifiers and random geographic offsets. The app uploads data every few hours to Amazon Web Services for storage.

Subjects with HF were enrolled in an ongoing HF study at the Veterans Affairs Medical Center and Emory University Hospital in Atlanta, USA. The study protocol was approved by the IRB (#00075867) at Emory University. The AMoSS app was installed on subject's phone after they visited the HF clinic. The subject could also opt to stop the collection of individual data types at any time. The app passively collected data while the clinical team recorded the clinical events, which consisted of hospital visits with compensated or decompensated status during the enrollment.

Table 2.2: AMoSS HF dataset description. If the metric is not available, the participant is excluded from that row.

| | |
|---|---|
| Num. comp. events | 62 |
| Num. decomp. events | 48 |
| Avg. comp. events per person | 2 |
| Avg. decomp. events per person | 2 |
| Avg. ejection fraction (%) | 35 |
| Gender | 93% male |
| Age (mean ± std) | 67 ± 8 |
| BMI (mean ± std) | 31 ± 6 |
| Employment | Employed: 3<br>Unemployed: 5<br>Retired: 7 |

There were 28 participants (26 males) who contributed at least one clinical event during their enrollment. To be included in the study, participants need to have a diagnosis consistent with congestive heart failure as noted in the electronic medical records within the Emory Health Network. Additionally, participants needed to be over the age of 18, able to consent to a clinical study, and spoke English as their primary language. Patients were ineligible for participation in the study if they had been diagnosed with a terminal illness with a life expectancy of less than six months or if they were enrolled in a hospice program. Additionally, participants could not have been enrolled in clinical study that precluded them from participating in another clinical study. Finally, participants had to be willing and able to comply with the use of their smartphone as indicated in the study. Table 2.2 shows more details about the participants in the dataset.

### 2.6.4 Duchenne muscular dystrophy dataset

Fifty five participants who were aged 5 to 17 years participated in the DMD study conducted at Children's Healthcare of Atlanta medical center. The study was approved by Emory University Institutional Review Board. Participants were recruited during regularly scheduled appointments. Informed consent was obtained from parents or legal guardians, and assent was obtained from each participant.

At the time of enrollment, 37 participants were taking steroids. Actigraphy devices were provided for 31 participants. Nine did not wear the device at the appropriate times or did not wear it at all. From the remaining 23 participants, 14 were ambulatory. During the study, each participant wore the Actiwatch 2 (AW2; Philips Respironics, Bend, Oregon) research watch for up to 10 days on their non-dominant wrist. The research watch recorded data continuously each day and summarized data into 30-second epochs. The median recording duration was seven days.

# CHAPTER 3
# CHANGE-POINT DETECTION FOR SLEEP STATE ESTIMATION

## 3.1 Rapid state change detection techniques

Non-stationarity is one of the key characteristics of human physiology and activity, driven by structured working days, alarms, unpredictable human interaction, etc., as well as intrinsic changes in the central nervous and cardiovascular systems. Many techniques have been proposed to artificially remove non-stationarities such as 'detrending', or removing a mean, slope, or nonlinear fit in an arbitrary piecewise manner [78, 79]. However, such approaches tend to create large artifacts around changes in stationarity [80, 81]. Moreover, many useful time series analysis techniques assume stationarity. Therefore, change point detection – the estimation of points in time where the probability distribution of a stochastic process changes – can enable the analysis of stationary segments of data and reveal underlying structure. For instance, Bernaola-Galvàn [82] used time series segmentation and change point detection to investigate non-stationaries in human HR time series and found mean level jumps between HR segments were smaller in heart failure patients compared to healthy controls. Furthermore, HR interval segments were found to follow a power law distribution, for both heart failure patients and healthy controls.

In this work, six change point detection methods were applied to a realistic artificial dataset, which consisted of beat-to-beat (RR) interval time series data generated by a previously published model [83], hereafter referred to as "RRGen". The performance of each change point detection algorithm was evaluated as a function of noise, tolerance (time between a true and estimated change point), and arrhythmia (ectopy). Using artificial data enables the assessment of algorithmic performance by comparing estimated change points against true change points. Moreover, if the artificial dataset is realistic, parameters of the change point detection method can be optimized for use on real data that exhibits similar statistics. By using the knowledge of the exact time of state transitions inherent in RRGen, we tuned parameters of each algorithm to detect change points during in real overnight recordings.

### 3.1.1 Artificial beat-to-beat interval time series data generation

The 'RRGen' algorithm was used to generate realistic 24-hour RR time series (tachograms) from a model of cardiovascular interactions and transitions between physiological states [83]. The model incorporates short-range variability due to Meyer waves and respiratory sinus arrhythmia, and long-range transitions in physiological states, by using switching distributions extracted from real data. The model incorporates both short and long-term variability, modeled on the normal sinus rhythm database [84]. Different tachograms can be produced by calling RRGen with different seeds and one example is illustrated in Figure 3.1.

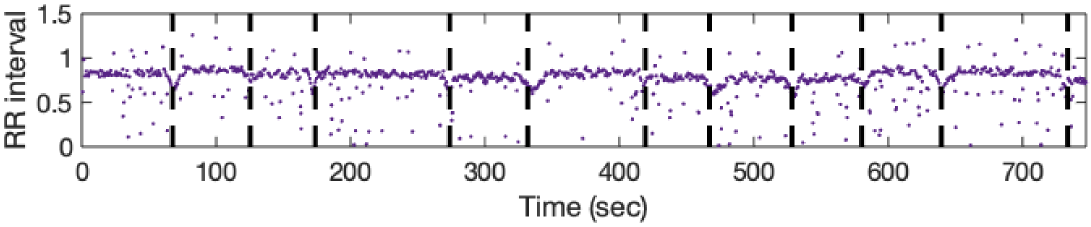


Figure 3.1: Artificial RR Interval data generated using RRGen. RR intervals are shown as purple points and the true change points are shown with dashed black lines.

### 3.1.2 Change point detection methods

*Recursive Mean Difference Maximization*

Recursive Mean Difference Maximization (RMDM) was proposed by Bernaola-Galvàn [82] to study scaling behavior of the human heart rate, and has been tested on artificially generated non-stationary time series with different statistical properties and real data [85, 86]. The method recursively maximizes the difference in the mean values between adjacent segments. Given the input signal $S = \{x_1, x_2, ..., x_N\}$ of length $N$, a sliding pointer is moved from the left to right, splitting the signal into $S_1 = \{x_1, x_2, ..., x_j\}$ and $S_2 = \{x_{j+1}, x_{j+2}, ..., x_N\}$ subsequences with $N_1$ and $N_2$ number of samples respectively, where $j$ is index at which the split occurs. Means of the subsequences $S_1$ and $S_2$ are calculated as:

$$\mu_1 = \frac{1}{N_1} \sum_{x_i \in S_1} x_i, \quad \mu_2 = \frac{1}{N_2} \sum_{x_i \in S_2} x_i. \tag{3.1}$$

The means are compared using the Student's $t$-statistic:

$$t(S_1, S_2) = \left| \frac{(\mu_1 - \mu_2)}{\sqrt{\sigma_P}} \right| \tag{3.2}$$

where $\sigma_P$ is the pooled variance, defined as

$$\sigma_P = \frac{(N_1 + N_2)(V(S_1) + V(S_2))}{((N_1 + N_2 - 2)N_1 N_2} \tag{3.3}$$

and $V(S)$ is the sum of squared deviations of the data in the signal $S$:

$$V(S) = \sum_{x_i \in S} (x_i - \mu)^2 \tag{3.4}$$

Student's $t$-statistic is calculated as a function of the index $j$ in the time series.

A candidate change point $j_{max}$ is selected, at which $t(j)$ reaches the maximum $t_{max}$.

The significance level $\mathcal{P}(\tau)$ of $j_{max}$ is calculated as $\mathcal{P}(\tau) = \{t_{max} \leq \tau\}$, where $\mathcal{P}(\tau)$ could not be obtained in a closed analytical form and was numerically approximated by Bernaola-Galvàn as

$$\mathcal{P}(\tau) = \left\{1 - I_{\left[\frac{\nu}{\nu+\tau^2}\right]}(\delta\nu, \delta)\right\}^{\gamma} \tag{3.5}$$

where $\gamma = 4.19 \ln N - 11.54$, $\delta = 0.40$, $N$ is the length of the signal, $\nu = N - 1$ is the number of degrees of freedom, and $I_x(a, b)$ is the incomplete beta function.

If $\mathcal{P}(\tau)$ exceeds a predefined threshold $P_0$ (0.95 as in this work), the signal is split into two subsequences. Before index $j$ is confirmed as a change point, $\tau$ between the two candidate subsequences is also calculated to see if the significance exceeds $P_0$. To reduce false positives, the condition that the minimum candidate segment length should be greater than $l_0$ was added. This procedure is repeated for each new sub-sequence until splitting the signal into candidate subsequences that differ by a significance level $P_0$ is not possible.

*Bayesian Blocks*

This algorithm was proposed by Scargle *et al.* and detects change points via dynamic programming [87]. The Bayesian Blocks (BBlocks) approach fits a piecewise constant signal model to data by maximizing a fitness measure specified according to the data type.

The total fitness of the partition $\mathcal{P}$ of the signal $S = \{x_1, x_2, ..., x_N\}$ of length $N$ is additive and defined as

$$F\left[\mathcal{P}(S)\right] = \sum_{k=1}^{N_s} f(S_k) \tag{3.6}$$

where $N_s$ is the number of segments and $f(S_k)$ is the fitness of the $k$th segment $S_k$ derived from maximizing the log likelihood of blocks given $M$ point measurements $x_i$, $i \in 1, ..., M$

$$f(S_k) = \frac{(\sum_M x_i)^2}{4 \sum_M \sigma_i^2} \tag{3.7}$$

where $x_i$ is the $i^{\text{th}}$ data point and $\sigma_i$ is the error variance of the data point measurement. Note that in our case, since a data point's measurement error was unknown and the signal was normalized, the error variance was taken as $\sigma = 1$.

The BBlocks algorithm starts with a sub-signal of only one data point $x_{i=1}$, wherein only one segmentation is possible. In each step, a new datum $x_{i=i+1}$ is added to the signal and can be considered the last point of the last segment of a possible optimal segmentation of samples. The starting point $r$ of the last segment of this optimal partition is obtained at each step by the following

$$r_i = \text{argmax}[f(r) + F\left[\mathcal{P}^{opt}(r-1)\right]] \tag{3.8}$$

where the definition of variables is the same as for the equations (Equation 3.6) and (Equation 3.7) and $\mathcal{P}^{opt}(r-1)$ is the optimal partition from the previous step.

When the last data sample is presented to the algorithm, the calculated value of $r_N$ becomes the last change point, marking the beginning of the last segment. This segment is removed, the data point before $r_N$ is considered the last datum, and the corresponding value of $r_i$ is assigned to the next change point. All change points are found by starting from the end, moving towards the start of the time series, and iteratively peeling off blocks.

### *Binary Segmentation*

Binary Segmentation (BiS) has been widely used in change-point detection analysis and is computationally fast ([88, 89]). The procedure minimizes a cost function, and starts by searching for a change-point $\tau$ in the input signal $S = \{x_1, x_2, ..., x_N\}$ that satisfies the condition

$$C_{S_{1:\tau}} + C_{S_{(\tau+1):N}} + \beta < C_{S_{1:N}} \tag{3.9}$$

where $C$ is a cost function and $\beta$ is a penalty term that reduces over-fitting. If the condition in (Equation 3.9) is met, $\tau$ becomes the first estimated change-point, and $S_{1:\tau}$ and $S_{(\tau+1):n}$ become the first subsequences. The process continues within these segments until data cannot be divided any further. Cost function in the above equation is given by

$$C_{S_{\tau_i:\tau_{i-1}}} = -2 \log L(\theta | S_{\tau_{i-1}:\tau_i}) \tag{3.10}$$

where $L$ is the likelihood function. If the data is assumed to follow the Normal distribution, log likelihood becomes

$$-2\log(L) = \sum_{j=\tau_i-1}^{\tau_i} \log(2\pi) + \log(\sigma_i^2) + \frac{x_i - \mu_i^2}{\sigma_i^2} \tag{3.11}$$

Then the cost function of mean changes in the time series can be expressed as

$$C_{S_{\tau_i:\tau_{i-1}}} = \sum_{j=\tau_i-1}^{\tau_i} \frac{(x_j - n_i^{-1} \sum_{j=\tau_i-1}^{\tau_i} x_i)}{\sigma^2} \tag{3.12}$$

Any cost function $C$ can be adapted for use in this framework, and one or more change-points can be detected.

*Pruned Exact Linear Time*

This algorithm was proposed by Killick *et al.* [89] and minimizes a cost function which chosen according to prior probability distribution of data. Algorithm calculates following minimization

$$\begin{aligned}
F(s) &= \min_{\tau \epsilon \tau_s}(\sum_{i=1}^{m+1}[C(S_{(\tau_{i-1}+1):\tau_i}) + \beta]) \\
&= \min_t(\min_{\tau \epsilon \tau_t}(\sum_{i=1}^{m+1}[C(S_{(\tau_{i-1}+1):\tau_i}) + \beta]) + C(S_{(t+1):n}) + \beta) \\
&= \min_t(F(t) + C(S_{(t+1):n}) + \beta)
\end{aligned} \tag{3.13}$$

Pruned Exact Linear Time (PELT) assumes that for all $t < m < T$, there is a constant K that satisfies

$$C(S_{(t+1):m}) + C(S_{(m+1):T}) + K \leq C(S_{(t+1):T}) \tag{3.14}$$

After each change point is estimated, pruning is performed by removing points that satisfy the condition

$$F(t) + C(S_{(t+1):m}) + K \geq F(m) \tag{3.15}$$

because these removed points cannot be the last optimal change point for $T > m$. In this work, the built-in MATLAB r2019a function ′`findchangepts.m`′ was used.

*Bayesian Online Change Point Detection*

Adams and MacKay proposed a Bayesian Online Change Point Detection (BOCPD) algorithm [90] and further developed by Turner *et al.* [91]. This method estimates change points using Bayesian inference, whereby the posterior probability of the time since the last change point, referred to as "run length", is calculated sequentially.

For normally distributed data with unknown mean and variance, the conjugate prior on observations follows a normal-inverse-gamma distribution with $\nu$, $\alpha$, and $\beta$ hyper parameters. On each step of the algorithm, a new datum $x_{i=i+1}$ is added to the analyzed signal and $x_t$ indicates data sample at time t. The posterior predictive probability of a new datum has the form of a non-standardized Student's t-distribution [92] with $2\alpha$ degrees of freedom, center at $\mu$, and PPV $\frac{\alpha\nu}{\beta(\nu+1)}$.

The posterior predictive probability of segment length $r$ at data point $i$ is given by

$$\pi_t^{(r)} = t_{2\alpha_t}(x_t|\mu_t, \frac{\beta_t(\nu_t+1)}{\alpha_t\nu_t}) \tag{3.16}$$

Using predictive probability, growth probabilities are calculated as

$$P(r_t = r_{t-1}+1, x_{1:t}) = P(r_{t-1}, x_{1:t-1})\pi_t^{(r)}(1 - H(r_{t-1}) \tag{3.17}$$

where $H(\tau)$ denotes the hazard function of a change point occurring. If intervals between change points are assumed to follow an exponential distribution with timescale $\lambda$, the hazard function becomes $H(\tau) = 1/\lambda$. The change point probability at time $t$ is

$$P(r_t = 0, x_{1:t}) = \sum_{r_{t-1}} P(r_{t-1}, x_{1:t-1})P(x_t|r_{t-1}, x_t^{(r)})H(r_{t-1}) \tag{3.18}$$

The distribution of run lengths is calculated as

$$P(r_t|x_{1:t}) = \frac{P(r_t, x_{1:t})}{\sum_{r_t} P(r_t, x_{1:t})}. \tag{3.19}$$

Finally, the hyperparameters are updated according to

$$\begin{aligned}
\mu_{t+1} &= \frac{\nu_t \mu_t + x_t}{\nu_t + 1}, \\
\nu_{t+1} &= \nu_t + 1, \\
\alpha_{t+1} &= \alpha_t + 0.5, \\
\beta_{t+1} &= \beta_t + \frac{\nu_t (x_t - \mu_i)^2}{2(\nu_t + 1)}
\end{aligned} \tag{3.20}$$

as defined in the conjugate Bayesian analysis of the Gaussian distribution [92]. This process repeats for the remaining data samples of the analyzed signal. After the run length distribution is calculated for all samples, indices with maximum probabilities are used to estimate the location of change points.

*Modified Bayesian Online Change Point Detection Algorithm*

Each new datum added to the analyzed signal results in one of two possible events for the segment length $r_t$: 1) it increases so $r_t = r_{t-1} + 1$, or 2) a change point occurs and $r_t = 0$.

Originally, for each value of $t$, the run length probability vector $v$ is sequentially calculated. $v(1) = P(r_t = 0)$ is calculated by (Equation 3.18). $v(2 : t)$ is the probability of each possible run length at time $t$ and calculated by (Equation 3.17). This vector is normalized by (Equation 3.19). Finally, all run length probability vectors are concatenated, and the run length with the maximum probability is selected as the run length for time $t$.

Although the run length should drop to zero as a change point is encountered in the time series, this approach fails to find $r_t = 0$. Due to sequentially calculating run length probability vectors, $v(1)$ is rarely the maximum row in the vector, so the run length is almost never set to zero.

A simple modification enables the correct selection of $r_t = 0$ when change

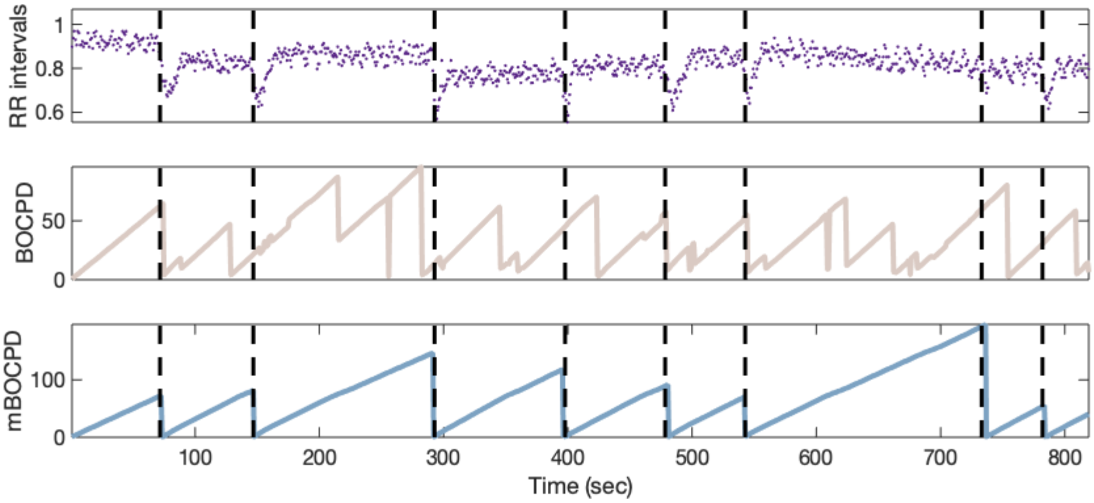


Figure 3.2: Top plot shows the artificially generated RR interval data, true change points are shown in black dashed lines. Middle plot shows the run length for BOCPD. Bottom plot shows the run length for mBOCPD. Note that the false triggering of BOCPD algorithm is avoided using the mBOCPD algorithm.

points are encountered. The growth probability, e.g. probability of continuing the current run, is calculated as

$$P(r_t = r_{t-1} + 1, x_{1:t}) = \sum_{r_{i-1}} P(r_{t-1}|x_{1:t-1}) P(x_t|r_{t-1}, x_t^{(r)})(1 - H(r_{t-1})) \qquad (3.21)$$

The growth probability is compared to the change point probability $P(r_t = 0|x_{1:t})$. If the change point probability is higher, the segment is ended and the run length becomes zero. Otherwise, the run length vector is increased by one. This approach is illustrated in Figure Figure 3.2 and is denoted as the Modified Bayesian Online Change Point Detection (mBOCPD) algorithm.

### 3.1.3 Performance metrics for change point detection

In order to assess how well estimated change points mapped to true change points, the following performance measures were used:

- **True Positive (TP)**: An estimated change point within a temporal tolerance $\gamma$ of a true change point was labeled as a true positive. If more than one estimated change point occurred within $\gamma$ of a true change point, only one estimated change point was counted towards the total number of true positives.

- **False Positive (FP)**: If the estimated change point was not within the tolerance, $\gamma$, of any true change point, then it was labeled as a false positive.
- **False Negative (FN)**: If there was not any estimated change point within the tolerance $\gamma$ of a true change point, then a false negative was recorded.

True negatives cannot be counted, since these would overwhelm any statistics and depend heavily on the sampling frequency. The recall or true positive rate ($TPR = TP/(TP + FN)$) and positive predictive value ($PPV = TP/(TP + FP)$) were calculated in this analysis. The true negative rate was not calculated because there were many more non-change points versus change points in the data. Instead, the number of false positives was counted. Finally, the F1 score — the harmonic mean of TPR and PPV ($F1 = 2TP/(2TP + FP + FN)$) - was used to optimize the parameters of each algorithm.

### 3.1.4 Parameter selection for change point detection algorithms

To find optimal parameters for each algorithm, the F1 score was maximized via grid search using realistic search ranges for each parameter. For RMDM, the minimum segment length ranged from $l_0 = \{6 : 1 : 12\}$. For BBlocks, the free parameter ranged from $\gamma = \{1.5 : 0.5 : 4, 5\}$. After $w_0$ and $p_0$ was set, the optimal posterior probability cut-off level was searched for in the range $\{0.5 : 0.1 : 1\}$. If the posterior probability of a point was higher than that level, it was labeled as a change point. In the analysis using the BOCPD method, the expected segment length was tested in the range $\lambda = \{100 : 100 : 2000\}$. For mBOCPD, the expected segment length ranged from $\lambda = \{10 : 10 : 100\}$ for RRGen. PELT was tested via detecting changes in root mean square level, standard deviation, mean, and "linear" mode – which finds the locations at which the mean and the slope of the signal change most abruptly. For BiS, changes in mean or mean and standard deviation were evaluated at the same time, and assumed data were drawn from normal distributions. For Hannan-Quinn, BIC, and AIC information criteria, the following penalties were tested respectively: $\beta = 2p \times \log(\log(N))$, $\beta = p \times \log(N)$, and $\beta = 2 \times p$. Parameters selected via grid search were used to analyze artificial RR interval data, for RMDM a minimum segment length $l_0 = 7$ was used. For BiS changes based on mean were evaluated. For PELT changes were based on root mean square level. For BBlocks a free parameter value $\gamma = 4$ was used. For BOCPD, $\lambda = 1840$ was used and for mBOCPD an expected segment length $\lambda = 80$ was used. Temporal tolerance $\gamma$ was set to 10 seconds.

### 3.1.5 Change point detection performance on artificial beat-to-beat interval time series

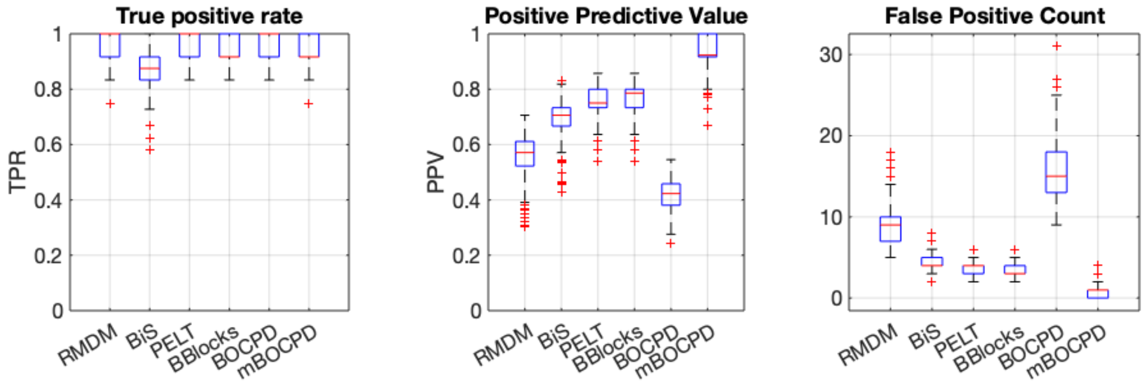


Figure 3.3: Performance of methods for detecting change points on the artificial data.

Using RRGen, 500 samples of RR interval data of length 1000 samples were generated. While RMDM achieved highest TPR when applied to artificial RR interval data, BOCPD had higher positive predictive value and fewer false positive counts compared to other methods as shown in Figure 3.3. When artificial

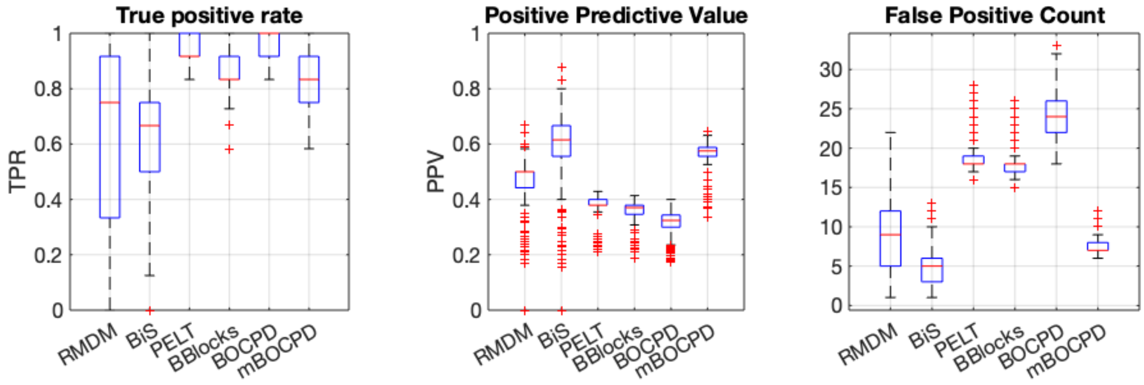


Figure 3.4: Performance of methods for detecting change points on the artificial data with noise probability set to 0.1.

noise was added (probability of noise = 0.01), BiS achieved the highest positive predictive value and the fewest false positive counts. The computation time ($t_c$) for methods was also recorded. For BOCPD and mBOCPD, $t_c$ was 0.178 and 0.164 seconds respectively. For RMDM $t_c = 0.159$, for BiS $t_c = 0.129$, for PELT $t_c = 0.004$, and for BBlocks $t_c = 0.023$. When the length of artificial RR interval time series was set to 9000 samples, the computational time was; for BOCPD $t_c = 7.149$, for BOCPD $t_c = 6.9146$, for RMDM $t_c = 1.629$, for BiS $t_c = 3.186$, for PELT $t_c = 0.010$, and for BBlocks $t_c = 0.426$.

## 3.2 A change point decoder for sleep-wake detection on memory-constrained wearables

The Change Point Decoder (CPD) uses the change points from wearable device signals to predict sleep or wake [93]. The method is inspired by the encoding/decoding framework in neuroscience [94], where a neural population response to a stimulus signal is observed in the form of spike trains. These responses are then used to train encoding models that describe the probability of the responses. When a spike train is observed from a group of cells, this model is used to ”decode” or estimate the stimulus signal. Similarly, in sleep/wake detection problems, the stimulus becomes the sleep/wake states, and alternations in these states result in the observed changes in PPG and actigraphy signals. Previous studies have shown that the mean and standard deviation of heart rate decreases during Non-REM sleep and increase during wakefulness [23, 24, 25]. We hypothesized that change point detection could be used to mark these alterations in the heart rate. Body movements have also been used as a sleep/wake identification feature in various studies over the years [13, 19, 20]. In this method, changes in the amplitude and gross body movements were detected to capture this information.

### 3.2.1 Preprocessing accelerometer and photoplethysmogram data

Initially, the Empatica E4 timestamp was synchronized with the PSG timestamp. A series of preprocessing steps were applied to PPG and accelerometer signals to convert these signals into a sequence of events. Firstly, the PPG signal was preprocessed using PhysioNet Cardiovascular Signal Toolbox [95]. Peak detection was performed using the *qppg* method provided with the toolbox, and the data was converted to peak-to-peak (PP) interval time series. Then, non-sinus intervals were detected and removed by measuring the change in the current PP interval from the previous PP interval and excluding intervals that change by more than %20. PP intervals outside of physiologically possible range were also removed to obtain NN interval time series, which was filtered using a Kalman filter to reduce noise [96, 97].

The accelerometer data was converted to activity counts following the approach by Borazio *et al.* [98]. Activity counts are the output format of most commercial actigraphy devices; data are summarized over 30-s epochs or time intervals. This conversion compresses information, and reduces required memory for storing data. Z-axis actigraphy data were filtered using a 0.25–11 Hz passband to eliminate extremely slow or fast movements [48]. The maximum values inside 1-s windows were summed for each 30-s epoch of data to obtain the activity count for each epoch. Figure 3.5 illustrates these preprocessing steps for accelerometer data.

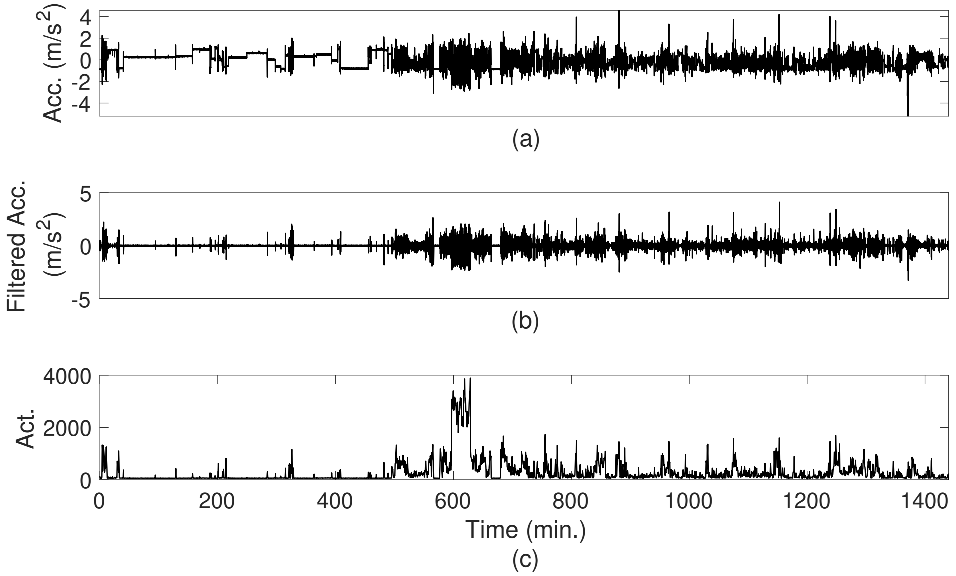


Figure 3.5: Preprocessing steps for the accelerometer data. Subplot (a) shows the raw accelerometer data from z axis and subplot(b) is the data after filtering. Activity counts obtained from summing the maximum values inside 1-second windows are shown in subplot (c).

Lastly, a tilt angle time series was derived from the raw accelerometer data to capture information that is not present in the activity count time series. Specifically, tilt angle, which is the angle between the gravitational vector measured by the accelerometer and the initial orientation with the gravitational field pointing downwards along the z-axis, can be calculated from the accelerometer reading as

$$\rho = \frac{a_z}{\sqrt{a_x^2 + a_y^2 + a_z^2}} \tag{3.22}$$

where $\rho$ is the tilt angle and $a_x$, $a_y$, and $a_z$ are the readings from x, y, and z axes of the accelerometer, respectively.

After obtaining the NN interval, actigraphy, and tilt angle time series, change-point detection techniques were applied to detect significant changes. We applied BiS technique to NN interval and actigraphy time series [89]. In a previous study, Yoneyama *et al.* stated that abdominal motion due to breathing causes 5° fluctuations, so 10° threshold is ideal for detecting body turnover events [99]. 10° changes in the tilt angle time series were stored as change events. In this way, all signals were represented as event sequences of the form $t_{1,i}, t_{2,i}, ..., t_{n,i}$ where $n \in \mathbb{Z}^+$ was the index of the change-point, $i \in \{1, 2, 3\}$ was the type of time series change-point occurred, and $t \in \mathbb{R}_{>0}$ denoted the time. Figure 3.6 illustrates the conversion to event streams.

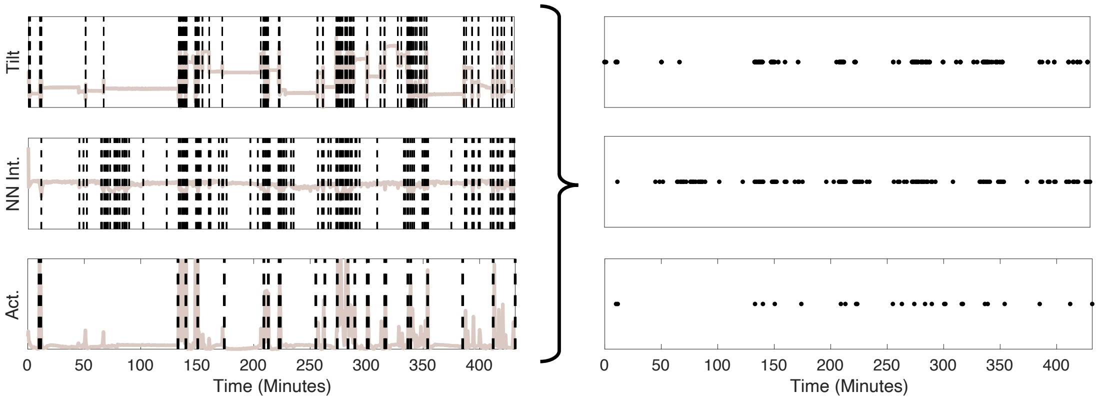


Figure 3.6: Conversion of NN interval, tilt, and actigraphy time series into point processes. Left hand side figures show the signals and detected change-points as dashed lines. Arrival times of each change-point $t_{n,i}$ are shown as dots in the right hand side figures.

### 3.2.2 Encoding step

In the encoding step, following the approach by Pillow *et al.* [94], change-points occurring in different signals were modelled. The intensity function $\lambda_{\text{NN}}(t)$ for NN interval time series was expressed as

$$\lambda_{\text{NN}}(t) = f(k \cdot x(t) + h \cdot z_{\text{NN,history}} + c_{NN,act} \cdot z_{\text{act}} + c_{NN,angle} \cdot z_{\text{angle}}) \tag{3.23}$$

where $x(t)$ was the sleep/wake stimulus that drives the changes in the signals. $k, h, c$ were stimulus, history, and coupling filters respectively. $z_{\text{NN,history}}$ represented the history of the Normal-To-Normal (NN) time series while $z_{\text{act}}$ and $z_{\text{angle}}$ were the windows of actigraphy and angle time series. $f$ was selected as the exponential function and it converted the summation into probability of spiking. The process generating the event streams was viewed as a Poisson Generalized Linear Model (GLM) and filter coefficients were estimated by fitting a GLM to the data. This generalized linear model approach allowed for both excitatory and inhibitory interactions between signals. We repeated same process for actigraphy and tilt angle time series to find the intensity functions $\lambda_{\text{act}}(t)$ and $\lambda_{\text{tilt}}(t)$ respectively. Figure 3.7 illustrates the encoding procedure for $\lambda_{\text{tilt}}(t)$.

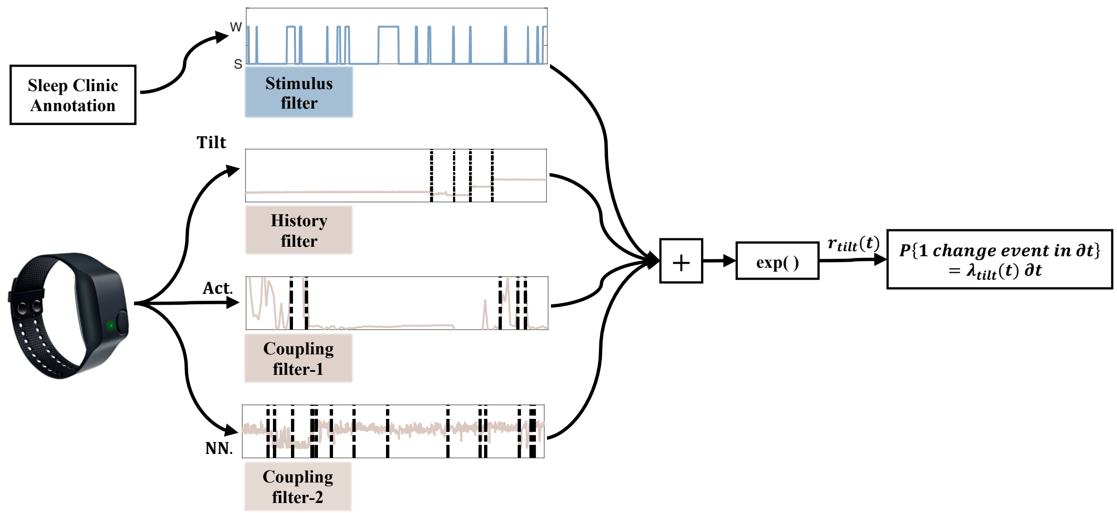


Figure 3.7: Encoding diagram for the angle time series. Raw signals from the Empatica devices are converted to change-point time series. History, coupling, and stimulus filters were applied on one-minute windows of data, summed and converted to spiking probability of tilt angle time series in time interval $\partial t$ using instantaneous firing rate $\lambda_{\text{tilt}}(t)$.

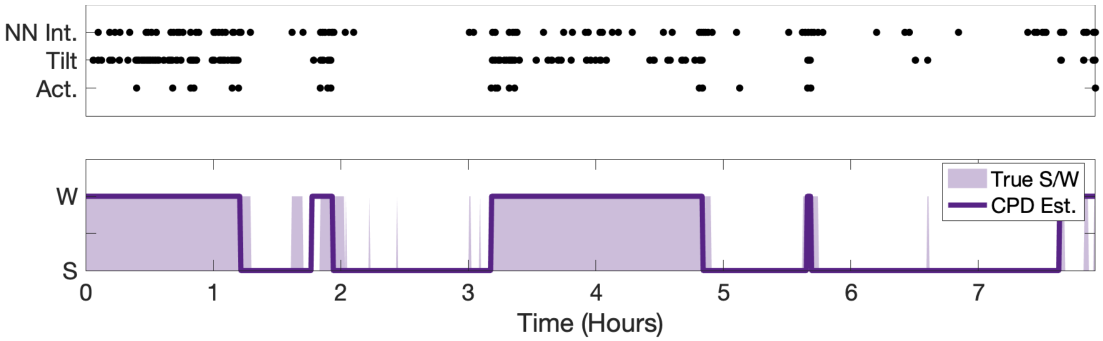


Figure 3.8: Decoding sleep/wake states from change event streams. Top plot illustrates the change events observed in NN interval, tilt and actigraphy time series. These change events are decoded back into sleep/wake states, as shown in the bottom plot.

### 3.2.3 Decoding step

In the decoding step, information about sleep/wake states was decoded back from the patterns of change observed in the signals, as shown in Figure 3.8. The log-likelihood function of events from a multidimensional process is given by

$$\log p(z|x,\theta) = \sum_{i,n} \log \lambda_i(t_{n,i}) - \sum_i \int_0^t \lambda_i(t)dt + const. \tag{3.24}$$

where $\theta = \{k, h, c\}$ represented model parameters from encoding step, $x$ was the sleep/wake stimulus, $z$ was the event streams. The posterior probability of the sleep/wake stimulus given the event streams was

$$\log p(x|z) = \log p(z|x) + \log p(x) + const. \tag{3.25}$$

Then, combining Equation 3.24 and Equation 3.25, penalized maximum likelihood estimate of the sleep/wake stimulus was calculated by minimizing

$$x_{est} = \underset{x}{\operatorname{argmin}}(-\log p(z|x) + \Lambda\|x_{\text{TV}}\|) \tag{3.26}$$

where $x_{est}$ was the estimate sleep/wake and $z$ was the observed change event streams. Total Variation (TV) norm prevented overfitting and preserved step-like properties of the sleep/wake stimulus. After estimation, the output $x_{\mathrm{est}}$ was thresholded and converted back to binary sleep/wake detection.

### 3.2.4 Hyperparameter selection for change-point decoder

The data window size for encoding model filters, the TV regularization parameter $\Lambda$, and the threshold were selected by sweeping a range of values and selecting parameters maximizing the F1 score on the training set. The F1 score was used to guide model selection since it is a combined metric for precision and recall. In this task, precision indicates how many epochs in the detected wake are correct, whereas recall refers to the percentage of total wake epochs correctly classified. Therefore F1 score, which combines precision and recall, proves to be a useful metric for this imbalanced classification scenario.

### 3.2.5 Performance on the Emory PTSD dataset

The Emory PTSD dataset was used for training and testing the proposed method since in this study, participants underwent PSG study and wore their research watch during the recording. The sleep technicians from the Emory Sleep Clinic annotated the recordings and these annotations became the ground truth in the analysis. The study population was assigned to four groups according to their Apnea-Hypopnea Index (AHI) and Periodic Limb Movement Index (PLMI) as follows:

- **Group 1:** Subjects with AHI $<$ 15 and PLMI $<$ 15
- **Group 2:** Subjects with AHI $\geq$ 15 and PLMI $<$ 15
- **Group 3:** Subjects with AHI $<$ 15 and PLMI $\geq$ 15
- **Group 4:** Subjects with AHI $\geq$ 15 and PLMI $\geq$ 15

All the data from the dataset were randomly split into two sets, with 70 subjects assigned to the training set and 32 subjects assigned to testing. Two-sample Kolmogorov tests were performed for age, Apnea–hypopnea index (AHI), periodic limb movement index (PLMI), and sleep efficiency of the subjects in the training and testing sets. Differences in these measures between the sets were not statistically significant, suggesting that the training set is representative of the testing set. Table 3.1 show ages and PSG-defined sleep efficiency in both sets.

Table 3.1: Participant Demographics and PSG Sleep Statistics in Training and Test Sets from Emory PTSD dataset. Mean (Standard Deviation) of variables in each group.

| | Training Set | | | Testing Set | | |
|---|---|---|---|---|---|---|
| Diagnosis | n | Age | PSG Sleep Efficiency | n | Age | PSG Sleep Efficiency % |
| Group 1 | 19 | 68.11 (2.31) | 75 (13) | 9 | 68.22 (2.64) | 74 (14) |
| Group 2 | 22 | 68.05 (2.03) | 72 (14) | 7 | 69.43 (1.62) | 74 (13) |
| Group 3 | 24 | 68.42 (2.87) | 74 (15) | 14 | 68.28 (1.68) | 67 (19) |
| Group 4 | 5 | 67.20 (3.42) | 65 (16) | 2 | 69 (0) | 74 (5) |
| All Subjects | 70 | 68.11 (2.48) | 73 (14) | 32 | 68.56 (1.93) | 71 (16) |

Hyperparameters selected on the training set for CPD are 1-minute window size, regularization parameter of 2, and threshold[3] of 0.22. Concordance between PSG and the method was evaluated on the testing set. The mean across subjects for total accuracy, sleep accuracy, wake accuracy, Cohen's Kappa, F1 score, Wake After Sleep Onset (WASO), and Sleep Efficiency (SE) were calculated. For both WASO and SE metrics, the error was calculated as the PSG gold standard minus estimated value. On the testing set, CPD achieved a total accuracy of 0.72, sleep accuracy of 0.70, and wake accuracy of 0.74. Kappa value was 0.40, which indicated fair agreement between the method and the gold standard PSG annotation. WASO and SE errors were 7.66 minutes and 2.09 respectively.

Table 3.2 shows the same experiment repeated by using each signal by itself, without the coupling filters between the different domains. Tilt angle signal model performed better than PPG and actigraphy models in terms of Kappa, F1 score, WASO error, and SE error performance metrics. However, all three single signal models resulted in lower total accuracy, Kappa, F1 score, and higher SE error when compared to the combined model with the coupling filters.

Table 3.2: Performances of single signal models in the testing set.

| | PPG model<br>Mean (SD) | Act. model<br>Mean (SD) | Tilt model<br>Mean (SD) |
|---|---|---|---|
| Total accuracy | 0.60 (0.14) | 0.69 (0.13) | 0.69 (0.15) |
| Sleep accuracy | 0.49 (0.19) | 0.89 (0.12) | 0.65 (0.22) |
| Wake accuracy | 0.83 (0.13) | 0.34 (0.17) | 0.75 (0.20) |
| Kappa | 0.25 (0.17) | 0.24 (0.19) | 0.36 (0.23) |
| F1 score | 0.60 (0.16) | 0.41 (0.17) | 0.61 (0.19) |
| WASO error (min) | -53.31 (89.39) | 65.13 (69.45) | -2.44 (65.21) |
| SE error (%) | 20.87 (22.37) | -16.54 (15.98) | 6.21 (18.13) |

*Comparison with other methods*

The Oakley (OA) sleep/wake detection method was also implemented using the same dataset. The algorithm is adapted for 30-second epochs following the approach by Kosmadopoulos *et al.* [100]. Actigraphy data is weighted and summed as follows

$$\begin{aligned} A_i = 0.04 \cdot E_{(i-4)} + 0.04 \cdot E_{(i-3)} + 0.2 \cdot E_{(i-2)} + 0.2 \cdot E_{(i-1)} + 2 \cdot E_{(i)} \\ + 0.2 \cdot E_{(i+1)} + 0.2 \cdot E_{(i+2)} + 0.04 \cdot E_{(i+3)} + 0.04 \cdot E_{(i+4)} \end{aligned} \tag{3.27}$$

where $i$ denotes the current epoch index and $E$ denotes the actigraphy count in the epoch. Then $A_i$ is compared to a predefined threshold to identify sleep/wake. In Actiwatch devices, there are three different thresholds: low (20), medium (40), and large (80). Since the wearable device is different in this study (Empatica E4,

Table 3.3: Sleep/wake identification performances in the Training Set for OA and CPD methods

| | Training Set | | | |
|---|---|---|---|---|
| | OA<br>Mean (SD) | 95% CI | CPD<br>Mean (SD) | 95% CI |
| Total Accuracy | 0.76 (0.10) | [0.73, 0.78] | 0.76 (0.12) | [0.73, 0.79] |
| Sleep Accuracy | 0.82 (0.15) | [0.79, 0.86] | 0.78 (0.19) | [0.73, 0.82] |
| Wake Accuracy | 0.61 (0.19) | [0.56, 0.65] | 0.72 (0.18)* | [0.67, 0.76] |
| Kappa | 0.41 (0.17) | [0.37, 0.45] | 0.46 (0.20) | [0.41, 0.50] |
| F1 Score | 0.59 (0.14) | [0.56, 0.63] | 0.64 (0.15)* | [0.60, 0.68] |
| WASO Error (min.) | -22.82 (69.04) | [-39.28, -6.36] | 13.17 (56.84)* | [-0.38 26.72] |
| SE Error (%) | 3.01 (15.74) | [-0.74, 6.76] | -1.59 (14.18)* | [ -4.97, 1.79] |

* Wilcoxon signed-rank comparison of two methods, 5% significance level. Abbreviations: CI, Confidence Interval; SD, Standard Deviation.

Empatica; Cambridge, MA), it could result in an actigraphy time series with a different amplitude range than Actiwatch and thresholds may not apply. Therefore, the threshold was selected using the training data to maximize F1 score and was set to 70.

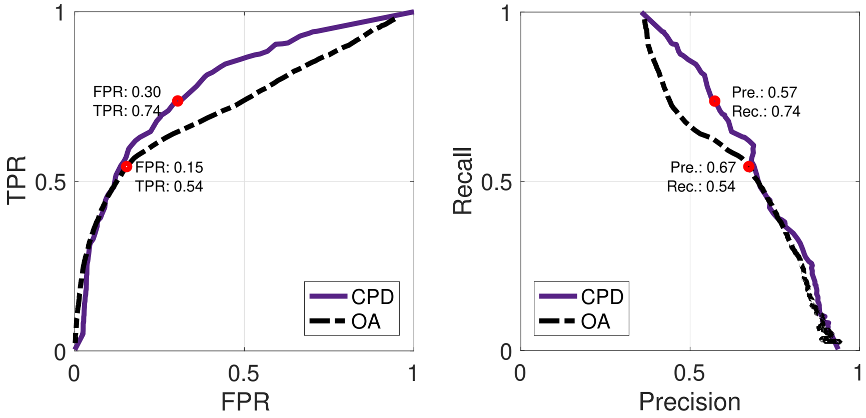


Figure 3.9: ROC and Precision-Recall curves for the CPD and OA methods. Performance of both methods is illustrated as their threshold varied. Operating points are shown with red circles on the plots.

Concordance between PSG and the two methods are evaluated on testing set. The mean across subjects for total accuracy, sleep accuracy, wake accuracy, Kappa, F1 score, WASO, and SE are shown in Table 3.3 and Table 3.4 for both methods.

Table 3.4: Sleep/wake identification performances in the Testing Set for OA and CPD methods

| | Testing Set | | | |
|---|---|---|---|---|
| | OA<br>Mean (SD) | 95% CI | CPD<br>Mean (SD) | 95% CI |
| Total Accuracy | 0.76 (0.09) | [0.72, 0.79] | 0.72 (0.14) | [0.67, 0.77] |
| Sleep Accuracy | 0.85 (0.12)* | [0.80, 0.89] | 0.70 (0.19) | [0.63, 0.76] |
| Wake Accuracy | 0.54 (0.20) | [0.47, 0.62] | 0.74 (0.21)* | [0.66, 0.81] |
| Kappa | 0.39 (0.17) | [0.33, 0.45] | 0.40 (0.24) | [0.31, 0.49] |
| F1 Score | 0.59 (0.14) | [0.54, 0.64] | 0.62 (0.20) | [0.55, 0.70] |
| WASO Error (min.) | -9.95 (63.75) | [ -32.94, 13.03] | 7.66 (67.34) | [-16.62, 31.94] |
| SE Error (%) | -0.03 (14.93) | [-5.42, 5.35] | 2.09 (16.81) | [-3.97, 8.15] |

* Wilcoxon signed-rank comparison of two methods, 5% significance level. Abbreviations: CI, Confidence Interval; SD, Standard Deviation.

Figure 3.9 illustrates ROC curve and Precision-Recall curve for both methods as their threshold is varied. Operating points selected using the training data are also marked with red circles on the plots. The AUC for the CPD method was found to be 0.78 and 0.67 for the OA method. Moreover, we observed from Figure 3.9 that it was possible to achieve similar performance to OA by changing the CPD method's threshold. However, it was not possible for OA method to reach the CPD's operating point by modifying the threshold value.

As shown in Table 3.3 and Table 3.4, the CPD method achieved greater accuracy for wake accuracy, Kappa, and F1 Score for both training and test sets. The difference between wake accuracy was statistically significant ($p < 0.05$) for the methods in both training and test sets. It can also be seen that OA overestimated WASO while wake accuracy is low. Note that the CPD method exhibited lower WASO error in both analyses. When using the medium threshold setting (40) is used for the OA method, total accuracy was 0.54, sleep accuracy was 0.38, and wake accuracy was 0.81 for the test set. The error in the number of sleep wake transitions in the test set was overestimated as -17.19 (36.13) for the OA algorithm and underestimated as 64.41 (34.80) for the CPD.

Figure 3.10 provides the Bland Altman analyses of the differences for SE and WASO for the OA and CPD methods for the Testing set. The modified Bland Altman plot shows that the Oakley method exhibited a bias towards overestimating WASO (see Figure 3.10, bottom left subplot). These plots also show that both methods exhibited similar performance as measured by SE error.

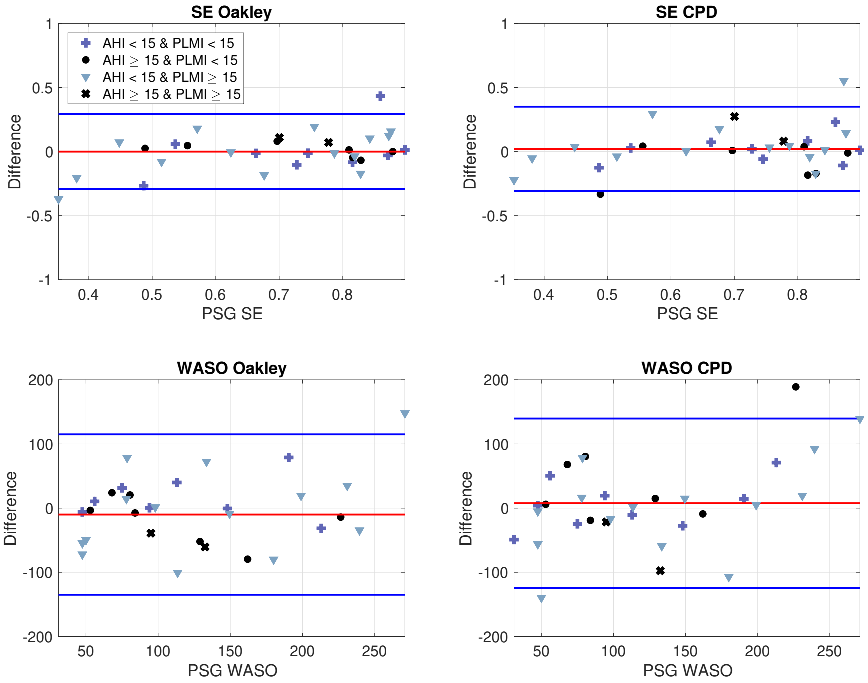


Figure 3.10: Modified Bland–Altman plots for sleep metrics in test set. X-axis shows ground truth (i.e. gold standard) PSG metrics and y-axis shows the difference between PSG and the estimates. Participants belonging to four subgroups determined by AHI and PLMI are indicated with different symbols.

Table 3.5 and Table 3.6 compare the results of both methods for all four groups in the test set. The CPD has a higher wake accuracy than the Oakley method in each subject group, while the Oakley method performs slightly better in terms of total accuracy.

Table 3.5: Sleep/wake identification performance in different disorder groups in the Testing set.

| | | Total Accuracy<br>Mean (SD) | Sleep Accuracy<br>Mean (SD) | Wake Accuracy<br>Mean (SD) | Kappa<br>Mean (SD) |
|---|---|---|---|---|---|
| Group 1 | Oakley | **0.78 (0.07)** | **0.88 (0.10)** | 0.52 (0.20) | 0.42 (0.13) |
| | CPD | 0.76 (0.07) | 0.76 (0.13) | **0.73 (0.23)** | **0.44 (0.20)** |
| Group 2 | Oakley | **0.78 (0.09)** | **0.84 (0.13)** | 0.55 (0.12) | **0.42 (0.10)** |
| | CPD | 0.75 (0.12) | 0.75 (0.13) | **0.63 (0.27)** | 0.37 (0.30) |
| Group 3 | Oakley | **0.73 (0.09)** | **0.84 (0.13)**[*] | 0.54 (0.25) | 0.35 (0.20) |
| | CPD | 0.69 (0.18) | 0.64 (0.23) | **0.78 (0.18)**[*] | **0.39 (0.25)** |
| Group 4 | Oakley | **0.75 (0.17)** | **0.78 (0.16)** | 0.67 (0.17) | **0.44 (0.33)** |
| | CPD | 0.70 (0.14) | 0.62 (0.21) | **0.87 (0.02)** | 0.42 (0.21) |

[*] Wilcoxon signed-rank comparison of two methods, 5% significance level.

Table 3.6: Sleep study statistic estimation performance in different disorder groups in the test set

| | | SE error<br>Mean (SD) | WASO error<br>Mean (SD) |
|---|---|---|---|
| Group 1 | Oakley | -0.07 (18.75) | -8.56 (78.67) |
| | CPD | 1.67 (10.89) | 5.33 (38.42) |
| Group 2 | Oakley | 0.71 (5.16) | -16.00 (37.62) |
| | CPD | -8.73 (14.38) | 47.21 (72.94) |
| Group 3 | Oakley | -1.68 (16.93) | -2.14 (69.85) |
| | CPD | 5.54 (19.42) | -1.04 (74.88) |
| Group 4 | Oakley | 9.02 (2.67) | -49.75 (15.20) |
| | CPD | 17.73 (13.65) | -59.50 (53.74) |

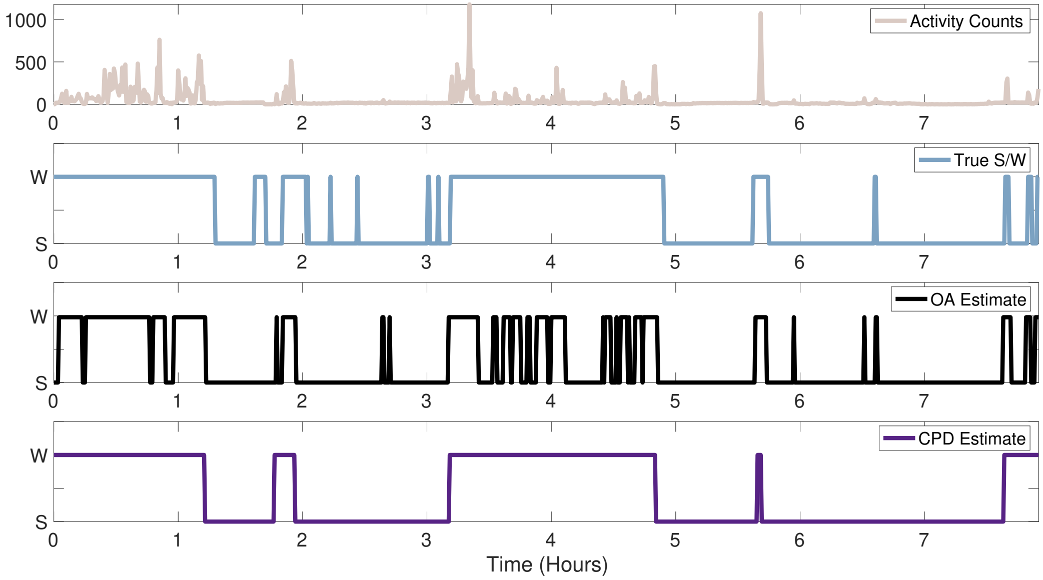


Figure 3.11: Comparison of OA and the CPD methods for one participant in the Emory PTSD dataset. First plot illustrates the activity counts from the Empatica research watch, second plot shows the ground truth annotations from the Emory sleep clinic. The third and fourth plots are the CPD and OA estimates respectively.

### 3.2.6 Performance on the data collected with Verily research watch

CPD method's performance was also tested on the Verily research watch. Using the research watches from AURORA study, 21 nights were recorded from 3 participants. In this internal dataset, participants wore research watch while undergoing a PSG study or wearing a Sleep Profiler (Advanced Brain Monitoring, Carlsbad, CA). Ground truth labels for each night's recording was obtained either from sleep technician annotations for the PSG or Sleep Profiler. The CPD model was not retrained due to the small size of this dataset but the weights from Emory PTSD dataset was directly used. With threshold set at 0.5 level and no adjustment, sleep accuracy was 0.73, wake accuracy was 0.77, F1 score was 0.31, and Cohen's kappa value was 0.22. If the threshold is readjusted on this dataset using the highest F1 score, sleep accuracy was 0.90, wake accuracy was 0.58, F1 score was 0.43, and Cohen's kappa value was 0.37. The data collection for this internal study is still ongoing and the performance after retraining the model entirely will also be investigated in the future work.

### 3.2.7 Discussion

This chapter of the thesis presents a novel method (CPD) for identifying sleep and wake states from movement and physiological signals collected using wearable devices. The method was comprised of three types of filters; stimulus, history, and coupling. Filter coefficients were estimated through a training process and then were used to detect sleep and wake states from change points. Our approach was flexible enough to incorporate various signal modalities and incorporating information from these results in higher wake detection performance. The CPD approach used a combination of movement-related and physiological signals, making it possible to overcome some of the limitations of previous algorithms based solely on actigraphy. For instance, the results demonstrate that the CPD method does not overestimate sleep and has high wake detection performance. Therefore, the CPD method can provide an unbiased solution to sleep/wake detection. The CPD modeled time series of discrete change events derived from wearable device signals and outputted a score of wakefulness which can be used to investigate gradual transitions between sleep and wake states within the epochs.

The OA method exhibited a higher sleep accuracy with respect to the CPD approach, which resulted in slightly higher total accuracy for OA since the prevalence of the sleep epochs in the data was relatively higher than the prevalence of wake epochs. By contrast, we observed a significant improvement in wake accuracy by using the CPD. Higher wake accuracy also resulted in lower WASO error for both training and test sets with the CPD. The OA method overestimated WASO and had lower wake detection accuracy, even though the threshold parameter was optimized during training (Table 3.3, Table 3.4). This outcome indicated that the Oakley algorithm misclassified sleep epochs as wake while being unable to recognize true wake epochs. A similar pattern was observed in subjects without any sleep disorder (Group 1) within the test set. This result could be due to the fact that when there is no movement, OA could not estimate that the subject was wake, as exemplified in Figure 3.11.

Periodic Limb Movement Disorder is characterized by episodes of limb movements during sleep, and these limb movements could bias the actigraphy based method into estimating a subject is awake. For Periodic Limb Movement Disorder subjects (Group 3), the CPD method had higher wake accuracy compared to OA, indicated in Table 3.5. However, this did not lead to significantly lowerWASO error due to the CPD method's lower sleep accuracy in this group, suggesting that limb movements had a similar effect in both methods.

Accurate estimates of WASO could become especially important in monitoring populations with difficulties falling or staying asleep. For example, WASO duration has been used as a diagnostic criterion for insomnia [101]. The OA method is known to have lower performance in detecting wakefulness for insomnia [21, 22]. In this study, optimizing the threshold parameter for OA did not yield a significant increase in wake accuracy. Therefore, the CPD method could be more useful in this population due to its higher accuracy in detecting wake epochs and the lower error in WASO. On the other hand, CPD method had a high error for estimating the number of sleep/wake transitions, which should be taken into account while applying the method on the insomnia population.

The proposed method only required the timestamps of the change points. Due to this fact, the CPD approach required less storage space than other methods. In this study, saving raw accelerometer and PPG signals for each subject resulted in 6.91 GB of data. However, if the change points alone were saved, stored data were only 1.3 MB. Using the CPD method reduced the required memory to 0.02% compared to other approaches that need the whole signal for feature extraction or training the models. As a result, the CPD method could result in immense memory (and energy) savings for large populations, applications with more data streams, and studies in which subjects are monitored over long periods.

This proposed method had some limitations. Since the signals were stored as change point time series and raw signals were not saved, the information in signal segments was lost. This could limit the data being used for other applications such as detecting or monitoring disorders like arrhythmia or sleep apnea. Also, it has been observed that the CPD approach has lower wake accuracy in subjects with sleep apnea (Group 2) compared to other groups. Future studies will explore adding a PPG-derived respiration signal to the model to improve performance in subjects with sleep apnea. A second limitation of the CPD is the lower number of sleep/wake transitions. The CPD method employs total variation regularization. While this regularization prevents overfitting and preserves piecewise constant structure of sleep/wake signal, it results in fewer switches between sleep and wakefulness.

### 3.2.8 Conclusion

In conclusion, this work presents the Change Point Decoder, which is a novel technique for sleep/wake identification in patients with highly disordered sleep. The CPD provides higher wake detection accuracy when compared to a solely actigraphy-based method. This superior performance could enable more accurate investigation of the vital role of awakenings during the night in various psychological disorders. The CPD method requires low memory in the wearable devices compared to existing methods, and therefore, it could prove beneficial in long-term studies. Moreover, as a method, the CPD has the ability to adapt to different and novel devices and signals beyond the accelerometer. In the following chapter, features related to biological rhythms will be investigated and CPD will be applied to data collected during daily life to estimate post-trauma outcomes from wearable data.

# CHAPTER 4

# CIRCADIAN RHYTHM DISRUPTION DETECTION

Acute diseases are health conditions that develop suddenly and have potential to improve with correct treatments. In this chapter, the extent to which outcomes developing post-trauma can be predicted from circadian rhythm changes was investigated, using longitudinal data passively collected from a research watch. Novel approaches to distinguishing between people that will and will not develop PTSD after exposure to a trauma is presented using the AURORA dataset, which was described in subsection 2.6.2. Figure 4.1 illustrates the AURORA study on the disease scale plot.

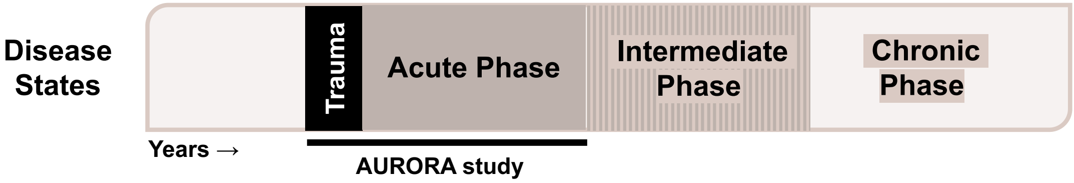


Figure 4.1: Illustration of AURORA study on disease time scales. The participants are enrolled into the study after exposure to a traumatic event and monitored through the acute phase using wearable devices.

The analysis in this chapter focuses on motion and heart rate data from the accelerometer and PPG sensors of the research watch. Figure 4.2 and Figure 4.3 show the data contribution of participants over the monitoring period. In these maps, if the participant did not share any samples from 1-hour window, it is marked as missing. It can be seen that missingness of the motion data increases gradually over time. On the contrary, heart rate data is missing from hour 10 to 20 consistently, possibly due to motion artifacts and low quality PPG data. For instance, there were only 44 participants with less than 25% missing heart rate data.

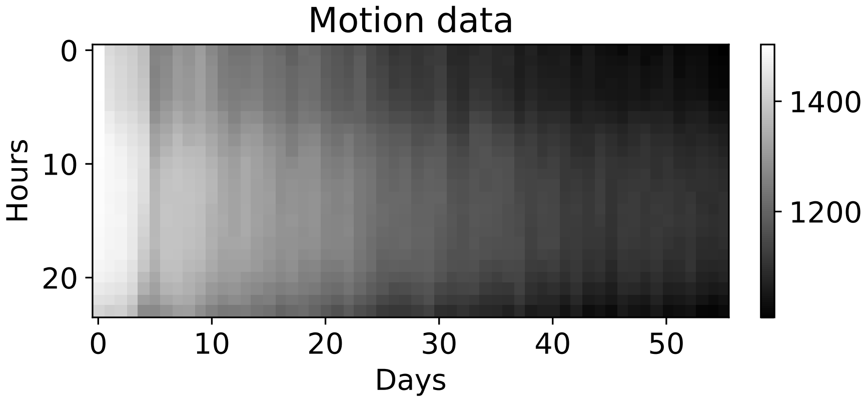


Figure 4.2: Binary missingness map of the motion data for the AURORA dataset. x axis is the 24 hours of the day, assuming the participant stayed in the same time zone from the Emergency Department enrollment and "0" hour is the midnight in this time zone. y axis of the map shows each day. If the participant did not share any samples from 1-hour window, it is marked as missing. Final map is obtained by summing how many participants shared data for each 1-hour window across 56 days.

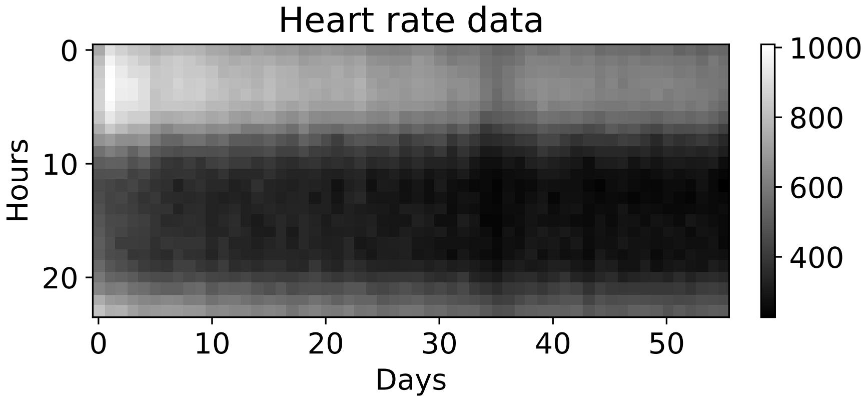


Figure 4.3: Binary missingness map of the heart rate data for the AURORA dataset. The map is constructed following the same procedure described in Figure 4.2 for the heart rate data.

## 4.1 Using circadian rhythm changes measured via wrist-worn research watch for post-trauma outcome classification

The first method is based on a classification algorithm fed by a set of motion features and HRV metrics. Data were analyzed using cosinor-based rhythmometry method [102] to completely automate the detection of rest/activity periods without the need for subjective information such as sleeping diaries or time zone information in the setting of both complete and missing data (the latter resulting from non-compliance or dead batteries). HRV metrics were extracted together with actigraphy features to quantify rest and activity states and examined the effect of varying the duration of data used to predict PTSD outcome.

### 4.1.1 Patient class labels and survey tools

Three clinical surveys were administrated in the ED – the Peritraumatic Distress Inventory (PDI), PTSD Checklist for DSM-5 (PCL-5), and Michigan Critical Events Perception Scale (MCEPS) [103, 104], as shown in Figure 4.4. PCL-5 administered at the ED solicited information on symptoms 30 days prior to the traumatic event. The raw scores of these surveys were used as features to the models to determine if prediction of the outcomes is feasible without using the research watch data. Hereinafter, these surveys will be referred to as $PDI_{ED}$, $PCL-5_{ED}$, and $MCEPS_{ED}$ to indicate they were administrated at ED.

Three clinical surveys administered at the eighth week of the study were used to create the binary outcome classes. These outcomes could potentially be used to identify subjects who require intervention to prevent or reduce the severity of PTSD. Firstly, the PCL-5 survey scores were used to capture PTSD symptoms outlined by Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria [30]. The score $PCL-5 = 31$ was used as the threshold, following the recommen-

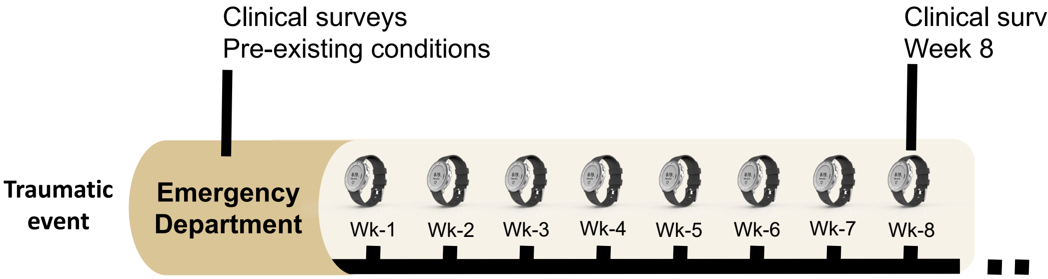
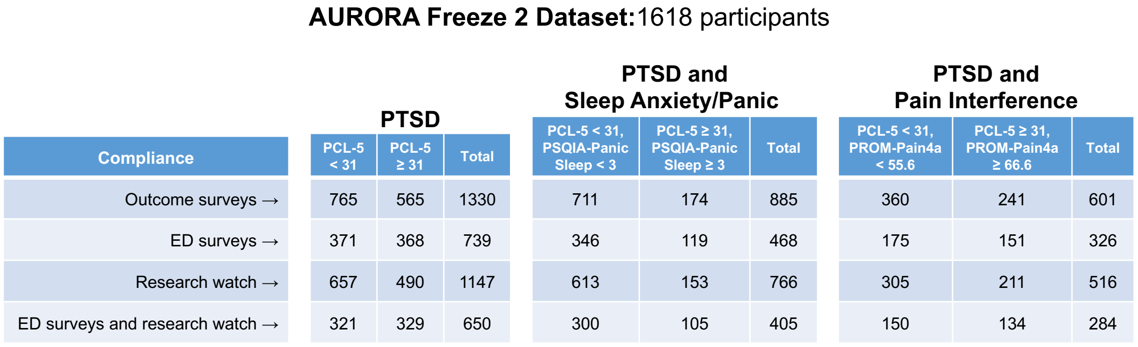


Figure 4.4: Timeline of data collection and clinical surveys in the AURORA study.

dation of the developers of the PCL-5 survey [105].

Secondly, since patients with PTSD report sleep disturbance, the PCL-5 questionnaire was combined with one item from Pittsburgh Sleep Quality Index Addendum (PSQIA) in order to measure sleep anxiety and panic [49, 106, 107, 108]. This item is referred to as PSQIA-PanicSleep. The question and response categories were modified as follows to assess the difficulty of staying asleep: *"In the 'reference period', how often did you awaken from sleep with severe anxiety or panic?" so that 0 = "never", 1 = "less than once a week", 2 = "1-2 nights a week", 3 = "3-4 nights a week" and 4 = "every or nearly every night"*. The cut-off for the survey was selected in order to separate participants with severe sleep disturbance. In this outcome, participants with $PSQIA - PanicSleep \geq 3$ and $PCL - 5 \geq 31$ were assigned to the first class while $PSQIA - PanicSleep < 3$ and $PCL - 5 < 31$ were assigned to the second class. This outcome is referred to as PTSD-Sleep Panic/Anx. outcome.

It has been shown in previous studies that chronic pain could accompany PTSD [32]. For the third outcome, the PCL-5 survey was combined with Patient-Reported Outcomes Measurement Information System (PROMIS) Pain Interference Short Form 4a (PROM-Pain4a) [109]. In this survey, the participant was asked to rate how much pain interfered with different areas of life on a 5-point scale (*1 = "not at all," 2 = "a little," 3 = "some," 4 = "a lot," and 5 = "extremely."*). The same scoring rules as the PROMIS Pain Interference Short Form 4a scale was used; the response values were summed and converted to a T-score. The T-score re-scales the raw score into a standardized score with a mean of 50 and a standard deviation of 10. A higher PROMIS T-score represents more of the concept being measured and the T-scores help in interpreting the PROMIS scores in a clinically meaningful way (More information about the T-scores could be found in www.healthmeasures.net). By using the PROMIS T-score guidelines, the cut-offs were selected for mild and severe pain interference following the guidelines for T-score interpretation. Participants with $PROM - Pain4a \geq 66.6$ (corresponding to a raw score of 16) and $PCL - 5 \geq 31$ were assigned to first class while $PROM - Pain4a < 55.6$ (corresponding to a raw score of 8) and $PCL - 5 < 31$ were assigned to second. This outcome is referred to as PTSD-Pain Int. outcome. Figure 4.5 illustrates the number of participants in each class, determined by week eight outcome surveys. Hereinafter, these surveys will be referred to as $PSQIA - PanicSleep_{week8}$, $PCL - 5_{week8}$, and $PROM - Pain4a_{week8}$ to indicate they were administrated at week eight.



**AURORA Freeze 2 Dataset:**1618 participants

| | PTSD | | | PTSD and Sleep Anxiety/Panic | | | PTSD and Pain Interference | | |
|---|---|---|---|---|---|---|---|---|---|
| Compliance | PCL-5 < 31 | PCL-5 ≥ 31 | Total | PCL-5 < 31, PSQIA-Panic Sleep < 3 | PCL-5 ≥ 31, PSQIA-Panic Sleep ≥ 3 | Total | PCL-5 < 31, PROM-Pain4a < 55.6 | PCL-5 ≥ 31, PROM-Pain4a ≥ 66.6 | Total |
| Outcome surveys → | 765 | 565 | 1330 | 711 | 174 | 885 | 360 | 241 | 601 |
| ED surveys → | 371 | 368 | 739 | 346 | 119 | 468 | 175 | 151 | 326 |
| Research watch → | 657 | 490 | 1147 | 613 | 153 | 766 | 305 | 211 | 516 |
| ED surveys and research watch → | 321 | 329 | 650 | 300 | 105 | 405 | 150 | 134 | 284 |

Figure 4.5: AURORA Freeze 2 Dataset overview and number of participants in each outcome group that is used in this research. Outcome surveys applied at week eight (PCL-5, PSQIA-PanicSleep, and PROM-Pain4a) were used to create the outcome groups. ED surveys included PDI, MCEPS and PCL-5 administrated at ED department following trauma. Top row of the tables indicates the number of participants that answered the outcome surveys, which is the maximum number available for the analysis. The rows below the first row indicate if the participants shared other modalities in addition to the outcome surveys.

### 4.1.2 Preprocessing steps for the accelerometer and photoplethysmogram data

The raw sensor data collected by the Verily research watch was preprocessed using the same procedure outlined in subsection 3.2.1. Movement data are commonly represented as a "double plot", which shows activity levels (measured via accelerometry in this case). Figure 4.6 illustrates this for one participant using eight weeks of actigraphy data. Each column is created by stacking two consecutive days of data. The first column shows activity levels on days 1-2, the second column shows days 2-3, and so on. Darker colors indicate lower levels of activity. If 24 hours are shown on the y-axis instead of 48 hours, we refer to this plot as movement map.

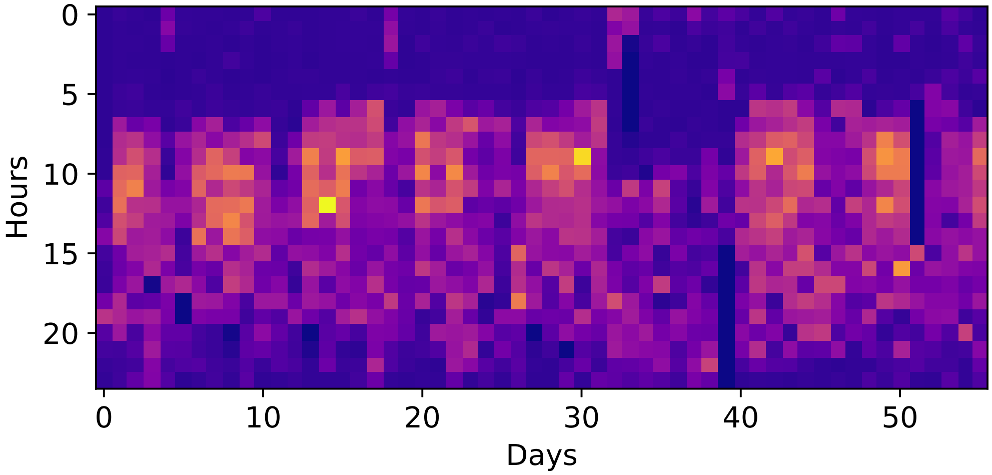


Figure 4.6: Detection of rest and activity regions from actigraphy data. Lighter colors indicate higher intensity movements. Deviations from the typical pattern are seen on days 30-40 in this example participant.

*Cosinor-based rest and activity region identification*

Single-component cosinor models were used to detect 24-hour rest and activity regions without any time-zone or sleep diary information [102]. Actigraphy data of each participant were split into 48-hour windows with an overlap of 24-hours.

The cosinor model with the following form was then fit to the data

$$Y(t) = M + K \cdot cos(2\pi t/\tau + \phi), \tag{4.1}$$

where $M$ is known as the *mesor*, $K$ is the *amplitude*, and $\phi$ is the *phase* of the circadian rhythm. By identifying the times at which the cosine fit crossed the mesor baseline, the start and end of rest and active segments of the day were determined. The process is illustrated in Figure 4.7.

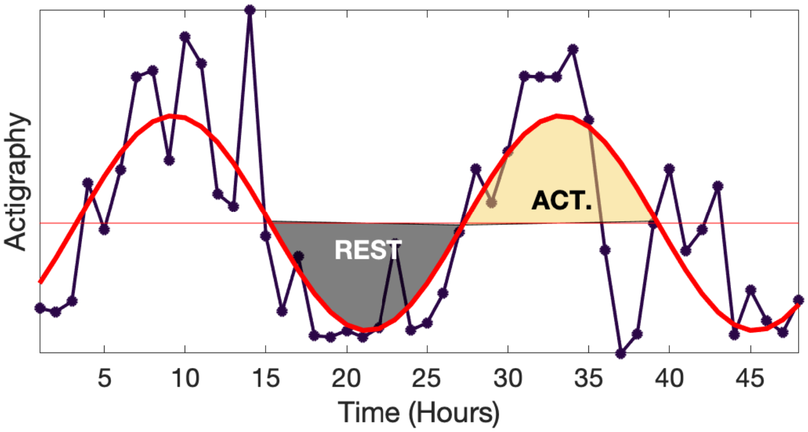


Figure 4.7: Fitting a cosine signal to actigraphy data to identify rest and activity regions. Detected rest and activity regions are marked on the figure with black and yellow shaded regions respectively.

### 4.1.3 Motion and heart rate variability feature sets

After preprocessing, the actigraphy signal in each 30-second epoch, together with the NN interval time series of each participant, was used for feature extraction. Table 4.1 and Table 4.2 describes the features extracted from these preprocessed signals. Features derived from the accelerometer signal included Interdaily Stability (IS), IV, the mean and standard deviation of movement in the detected rest and activity regions, circadian rhythm strength, rest start index, and cosinor-based rhythmometry metrics (Mesor, Amplitude, Phase).

Table 4.1: Movement feature set, derived from the accelerometer data from the Verily research watch.

| Feature Name | Abbr. | Description |
|---|---|---|
| Interdaily Stability | $IS_{rest}$, $IS_{act}$ | The ratio between the variance of the average actigraphy pattern around the mean and the overall variance. |
| Intradaily Variability | $IV_{rest}$, $IV_{act}$ | The ratio of the mean squares of the difference between all successive hours of actigraphy and the mean squares around the grand mean. |
| Movement | $MV_{rest,\mu}$, $MV_{rest,\sigma}$,<br>$MV_{act,\mu}$, $MV_{act,\sigma}$ | Mean and standard deviation of movement in the rest and active parts of the day. |
| Circadian Rhythm Strength | $CRS_{\mu}$, $CRS_{\sigma}$ | Average movement in active part of the day divided by average movement in rest region. |
| Rest Start Index | $RSI_{\mu}$, $RSI_{\sigma}$ | Hour index of the start of rest region. |
| Cosinor-based Rhythmometry | $Mesor_{\mu}$, $Mesor_{\sigma}$,<br>$Amplitude_{\mu}$, $Amplitude_{\sigma}$,<br>$Phase_{\mu}$, $Phase_{\sigma}$ | Cosine fit to the data (described in Equation 4.1), $M$ is the *mesor*, $K$ is the *amplitude*, and $\phi$ is the *phase* of the circadian rhythm. |

*rest*, *act*: Label indicating feature calculated during rest or activity periods
$\mu, \sigma$: Mean and standard deviation across the days analyzed

Table 4.2: Heart rate variability feature set, derived from the photoplethysmogram data from the Verily research watch.

| Feature Name | Abbr. | Description |
|---|---|---|
| Avg. NN Intervals | $NNmean_{\mu}$, $NNmean_{\sigma}$ | Mean of Normal-to-Normal (NN) intervals |
| IQR of NN Intervals | $NNiqr_{\mu}$, $NNiqr_{\sigma}$ | IQR of NN intervals |
| Kurtosis of NN Intervals | $NNkurt_{\mu}$, $NNkurt_{\sigma}$ | Kurtosis of NN intervals |
| Skewness of NN Intervals | $NNskew_{\mu}$, $NNskew_{\sigma}$ | Skewness of NN intervals |
| SDNN | $SDNN_{\mu}$, $SDNN_{\sigma}$ | Standard deviation of NN intervals |
| RMSSD | $RMSSD_{\mu}$, $RMSSD_{\sigma}$ | Root-mean square differences of successive NN intervals. |
| pNN50 | $pNN50_{\mu}$, $pNN50_{\sigma}$ | Mean number of times per hour in which the change in consecutive NN intervals exceeds 50 milliseconds. |
| Acceleration and Deceleration Capacity of Heart | $AC_{\mu}$, $AC_{\sigma}$,<br>$DC_{\mu}$, $DC_{\sigma}$ | Acceleration and deceleration capacity of heart, calculated with Phase-rectified signal averaging method. |
| High Frequency Spectral Content | $HF_{\mu}$, $HF_{\sigma}$ | Spectral content in high frequency band ($0.15Hz \leq LF \leq 0.4Hz$) |
| Low Frequency Spectral Content | $LF_{\mu}$, $LF_{\sigma}$ | Spectral content in low frequency band ($0.04Hz \leq LF \leq 0.15Hz$) |
| Total Power | $ttlpwr_{\mu}$, $ttlpwr_{\sigma}$ | Sum of energy in all bands. |
| LF/HF Ratio | $LFHF_{\mu}$, $LFHF_{\sigma}$ | Ration of LF and HF power. |
| Sample Entropy | $SampEn_{\mu}$, $SampEn_{\sigma}$ | Sample entropy of NN intervals |
| Approximate Entropy of Heart Rate | $ApEn_{\mu}$, $ApEn_{\sigma}$ | Approximate entropy of NN intervals |
| Signal Quality Index | $avgSQI_{\mu}$, $avgSQI_{\sigma}$ | Morphology-based quality of each beat |

$\mu, \sigma$: Mean and standard deviation across the days analyzed

A non-parametric way of quantitatively describing the rhythm is the Rest-Activity metrics, which include IS and IV [110]. IS can be defined as the ratio of the variance the 24 hour activity pattern around the mean and overall variance

$$IS = \frac{n \times \sum_{h=1}^{24} (\bar{a_h} - \bar{a})^2}{p \times \sum_{h=1}^{n} (a_i - \bar{a})^2}, \tag{4.2}$$

where $\bar{a}$ is the mean of all data, $\bar{a_h}$ are the hourly means, $a_i$ are the data points, and n is the number of data points in the interval. $p$ is set to 24 for the circadian rhythm assessment. IS quantifies the instability of the rhythm and the coupling to the environmental zeitgebers such as light and temperature [111].

IV is the ratio of the first derivative to overall variance and quantifies fragmentation of the rest-activity rhythm as

$$IV = \frac{n \times \sum_{i=2}^{n} (a_i - a_{i-1})^2}{(n-1) \times \sum_{h=1}^{n} (a_i - \bar{a})^2}. \tag{4.3}$$

IV is a measure of the fragmentation of the rest-activity rhythm and it quantifies the frequency of transitions between rest and activity. IV metric is higher if the participant is taking naps frequently during the day or have frequent awakenings during the night.

Rest-activity metrics have been used in various studies to assess the uncoupling of the rhythm to zeitgebers. In a previous work, we have derived the metrics from smartphone data to estimate heart failure decompensation related symptoms which include worsening fatigue and edema [64]. Huang *et al.* used the metrics to assess the age associated differences in circadian rhythm in a dataset of 65 young, middle-aged, old and the oldest age group participants [112]. The authors found that IV is highest in the oldest age group, which could indicate sleep disturbances. Figure 4.8 illustrates one application of rest-activity metrics to actigraphy data. It can be seen that a higher variation between each day results in a lower IS.

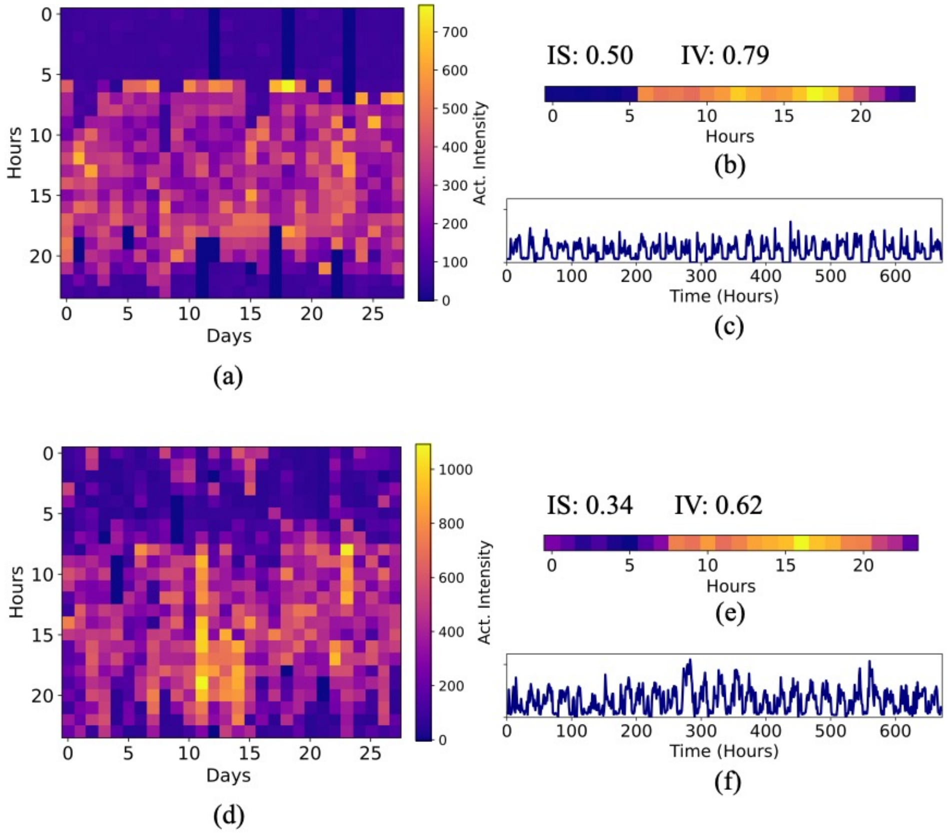


Figure 4.8: Movement patterns and rest-activity metrics for two participants. Subplot (a) is the double plot of the first participant, subplot (b) is the average daily activity, and subplot (c) is the plot of actigraphy data over hours. Subplots (d), (e), (f) are the double plot, average daily activity, and actigraphy plot for the second participant respectively.

Cosinor-based rhythmometry metrics can provide information about the participants' circadian rest-activity cycle [102]. To extract cosinor rhythmometry features, a cosine model described in Equation 4.1 was fit to the data. $M$ was mesor (baseline activity of subject), $K$ was amplitude (how active subject is during the day versus night), and $\phi$ was acrophase (a metric of the circadian cycle).

In addition, average movement in the active part of the day was divided by the average movement in the rest region to obtain the Circadian Rhythm Strength (CRS) metric. Active and rest parts of the day were determined by using the approach described in subsubsection 4.1.2. Rest Start Index (RSI) was the hour index of rest start region and approximated the start of the sleep period of the day. Lastly, mean and the standard deviation of the movement data from the rest active parts of the day was calculated (MV).

CRS, RSI, and cosinor-based rhythmometry metrics were extracted from each day within the window and then the mean and standard deviation of the metric were calculated. Similarly, for MV metrics, the mean and standard deviation of 12-hour rest and activity regions across the window were extracted.

The HRV feature set was derived using PhysioNet Cardiovascular Signal Toolbox and included time domain, frequency domain and entropy metrics [95]. More details about the HRV features used can be found in Table 4.2. Time domain features included $NN_{mean}$, $NN_{iqr}$, $NN_{kurt}$, $NN_{skew}$, $SDNN$, $RMSSD$, and $pNN50$. $NN_{mean}$ was the average heart rate without abnormal beats or arrhytmias. $NN_{iqr}$ was the statistical dispersion of the NN intervals. $NN_{kurt}$ measured how peaked or flat the distribution of the NN intervals was relative to a Gaussian. $NN_{skew}$ quantified the asymmetry of NN interval distribution. Strong asymmetries could be driven by rapid accelerations. $SDNN$ reflected the cyclic components that were responsible for variability. $RMSSD$ was associated with short-term rapid changes in the heart rate, and was correlated with vagus-mediated components of HRV. $pNN50$ assessed parasympathetic (vagal) activity. Frequency domain HRV metrics included $HF$, $LF$, $ttlpwr$, and $LFHF$. $HF$ reflected the modulation of vagal tone, primarily by breathing. $LF$ reflected modulation of sympathetic or parasympathetic tone by baroreflex activity. $ttlpwr$ was the sum of energy in all bands and was equivalent to variance. Lastly, $LFHF$ was the ration of $LF$ and $HF$ power and was indicator of sympathovagal balance. Entropy metrics included $SampEn$ and $ApEn$. $SampEn$ quantified the likelihood that two sequences similar for m points remain similar at the next point (i.e. match within a tolerance of r), not taking into account self-matches. $ApEn$ quantified the amount of regularity and the unpredictability of fluctuations over time-series data. In addition to these, signal quality of the PPG signal was calculated and used as an additional feature.

All HRV metrics were calculated in 5-minute segments with a 30-second overlap using the toolbox. Then 5-minute segments from the rest regions, detected by the cosinor method were selected in order to obtain the segments with the fewest movement artifacts and highest signal quality. The mean and standard deviation across the windows were calculated and used as features. Feature extraction was performed on a virtual computer in AWS, (48 vCPUs, 3.6 GHz, 96 GiB memory) and it took about three days for processing monthly data ( 700 participant's data on average).

### 4.1.4 Data organization for model training

As the first step in the pipeline, the data were adjusted by randomly undersampling the majority class in order to address the problem of class imbalance. This imbalance can be seen for PTSD-Sleep Panic/Anx. outcome, where the number of participants was 153 for the first class ($PSQIA - PanicSleep_{week8} \geq 3$ and $PCL - 5_{week8} \geq 31$) and was 613 for the second class ($PSQIA - PanicSleep_{week8} < 3$ and $PCL - 5_{week8} < 31$). Specifically, all participants from the minority class were used, and the same number of participants from majority class were randomly selected to obtain balanced classes. Under-sampling of majority class subjects was repeated in an external cross-validation fold, where n1 was defined as the number of majority class participants and n2 was the number of minority class participants. The external repeats were implemented n1/n2 times, and this ratio was rounded to the nearest integer.

### 4.1.5 Machine learning models

The mapping of the data or the normalized features (min-max normalization using [113]) into outcome classes is a supervised binary classification problem. All the models were written in the Python 3 language and the programming code is based on Scikit-learn [114]. Three different binary classifiers were trained for each experiment category as follows:

- SVM: An SVM is a supervised model that is designed to find the optimal separating hyper-plane with the maximum margin within the classes. Linear and radial basis function kernels were used.
- Logistic Regression: A logistic regression classifier uses a logistic function to model the probabilities of the outcomes. L2 regularization was used with the logistic regression classifier to achieve a robust model, minimize overfitting and reduce any effect of codependences without reducing the number of features. The regularization strength was set to the default level (1) of the Scikit-Learn logistic regression classifier.
- Multilayer Perceptron (MLP): A MLP is a type of supervised classifier with a feed-forward architecture, with one or more hidden layers between input and the output. A one-layer MLP with 100 neurons and L2 regularization was used, and these parameters were set at the default values for the Scikit classifier.

A five-fold cross-validation procedure was used for parameter tuning and model assessment and the class prevalence was adjusted to be identical in each fold. The model was trained on the data from all participants except one held-out fold, and the participants in the remaining fold were then used as the test data. This process was repeated to ensure testing on all participants. Performance metrics were calculated for each test fold, and the mean and standard deviation of each metric were calculated across the five folds. After extracting the features, the training phase of the classifier took an average of 0.57 seconds on a 2.3 GHz i5 Intel chipset.

### 4.1.6 Overview of experiments

Three categories of experiments were performed as follows and all models were tested for the full dataset and for the subset of participants whose PTSD outcome at week eight is different from baseline PTSD status assessed in the ED (ex: $PCL-5_{ED} < 31$ and $PCL-5_{week8} \geq 31$):

- **Experiment 1 (survey model)**: Prediction of eight-week outcome from ED survey data. The $PCL-5_{ED}$ solicits information on symptoms 30 days prior to the traumatic event. The raw scores of these surveys were used as features to the models.

- **Experiment 2 (research watch model)**: Prediction of eight-week outcome from the data. HRV and actigraphy features described in the previous sections were combined to obtain a feature matrix of 50 columns and models were trained to predict or classify the single corresponding eight-week outcome:

  - Using all participants and using a 56-day window.
  - Prediction of eight-week outcome using 7, 14, . . . , 56 days of HRV and actigraphy features, using participants who contributed data on all days. When an analysis window shorter than 49 days was used, the classifier was "predicting" the outcome at day 56 "ahead-of-time". However, when the analysis window size was 56 days for example, it reduced to a "classification" task.

  - Analysis of feature trajectories (daily averages of each feature in the 56-day window): Participants who report as non-PTSD ($PCL - 5_{ED}$ ¡ 31) in ED were isolated. Two subgroups were then created by looking at week eight surveys; participants who develop new-onset PTSD and those who remain non-PTSD. Then, the significance of each feature for these subgroups was tested using the Wilcoxon rank sum test.

- **Experiment 3 (fusion model)**: Fusion of research watch and survey models by concatenating the feature sets. Experiment 3 was implemented on participants who contributed both the research watch data and the ED survey data. Survey model and research watch model from previous experiments were also trained on this subset of participants to ensure results are directly comparable and the contribution of fusing modalities could be tested accurately.

### 4.1.7 Experimental results

*Results of Experiment 1*

The cross-validation performance of different types of classification models using ED survey-based features is shown in Table 4.3. Logistic regression classifier has achieved the highest AUC for all outcome types. Table 4.4 shows all metrics including accuracy, TPR, TNR, and PPV for the logistic regression classifier. Models showed high performance for all outcome types; 0.67, 0.70, and 0.70 accuracies for PTSD, PTSD-Sleep Anx./Panic, and PTSD-Pain Int. outcomes respectively.

The performance was evaluated for the participants for whom PTSD outcome changed from admission to week eight (N=270 for PTSD outcome, N=150 for PTSD-Sleep Anx./Panic outcome, N=110 for PTSD-Pain Int. outcome) without retraining the model. For these subsets of the participants, accuracies of 0.33,

Table 4.3: AUC comparison of different classifiers using ED surveys as features for eight-week outcome prediction. Results are reported as mean (standard deviation).

| **Outcome** | **Log. Reg** | **MLP** | **RBF SVM** | **Linear SVM** |
|---|---|---|---|---|
| PTSD | 0.73 (0.03) | 0.73 (0.03) | 0.72 (0.03) | 0.73 (0.03) |
| PTSD, Sleep Anx./Panic | 0.79 (0.04) | 0.79 (0.05) | 0.76 (0.07) | 0.78 (0.07) |
| PTSD, Pain Int. | 0.77 (0.04) | 0.77 (0.04) | 0.74 (0.04) | 0.76 (0.04) |

Table 4.4: Performance of logistic regression model using ED surveys as features for eight-week outcome prediction (N=739 for PTSD outcome analysis, N=468 for PTSD-Sleep Anx./Panic outcome analysis, N=326 for PTSD-Pain Int. outcome analysis). Results are reported as mean (standard deviation).

| **Outcome** | **Acc.** | **AUC** | **TPR** | **TNR** | **PPV** |
|---|---|---|---|---|---|
| PTSD | 0.67 (0.01) | 0.73 (0.03) | 0.64 (0.05) | 0.70 (0.05) | 0.69 (0.06) |
| PTSD, Sleep Anx./Panic | 0.70 (0.06) | 0.79 (0.04) | 0.67 (0.11) | 0.74 (0.07) | 0.72 (0.07) |
| PTSD, Pain Int. | 0.70 (0.04) | 0.77 (0.04) | 0.68 (0.03) | 0.73 (0.09) | 0.72 (0.09) |

0.32, and 0.34 was achieved for PTSD, PTSD-Sleep Anx./Panic, and PTSD-Pain Int. outcomes respectively.

*Results of Experiment 2*

Table 4.5 shows the performance of different classifiers when HRV and actigraphy features were used. It can be seen that similar to Exp. 1, logistic regression classifier performed the best for all outcome types. Models achieved the highest AUC of 0.70 and accuracy of 0.65 when the outcome is PTSD-Pain Int. However, the performance was lower for other outcome types; accuracy was 0.56 for PTSD outcome and 0.58 for PTSD-Sleep Anx./Panic. Table 4.6 shows the logistic regression classifier performance in detail for the research watch models. The model performance was similar for participants undergoing a change in the clinical status. The accuracies were 0.55, 0.59, 0.64 for PTSD, PTSD-Sleep Anx./Panic, and PTSD-Pain Int. outcomes respectively for this subset. For each outcome type, Figure 4.9 shows the feature importance determined by the absolute value of the logistic regression coefficients, averaged over folds. Figure 4.10 illustrates the AUC from each window size when participants with data contribution from all 56 days are considered. The best performance was achieved when all 56 days were used as the analysis window.

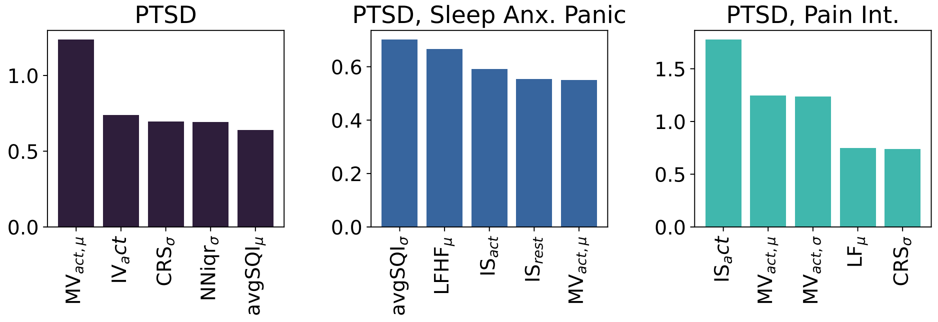


Figure 4.9: Feature importance for logistic regression models (window size=56 days). Highest five average absolute feature coefficients across folds are illustrated for each outcome

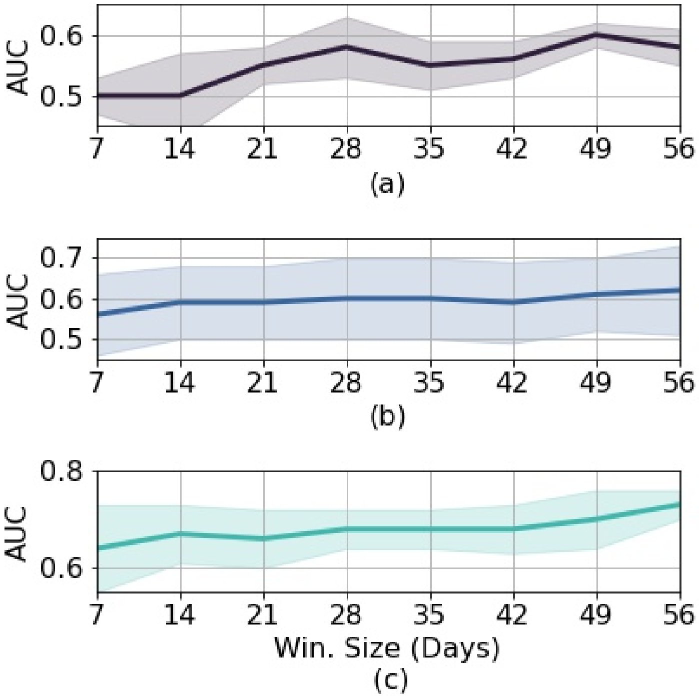


Figure 4.10: AUC of the logistic regression models with different window size selection. Subplot (a) shows the AUC for the PTSD outcome, subplot (b) shows PTSD-Panic Sleep/Anx. outcome, and subplot (c) shows PTSD-Pain. Int. outcome over the days

Table 4.5: AUC comparison of different classifiers using HRV and actigraphy features for eight-week outcome prediction. Results are reported as mean (standard deviation).

| **Outcome** | **Log. Reg** | **MLP** | **RBF SVM** | **Linear SVM** |
|---|---|---|---|---|
| PTSD | 0.56 (0.05) | 0.55 (0.04) | 0.54 (0.03) | 0.56 (0.05) |
| PTSD, Sleep Anx./Panic | 0.61 (0.06) | 0.60 (0.06) | 0.61 (0.07) | 0.59 (0.06) |
| PTSD, Pain Int. | 0.70 (0.02) | 0.69 (0.04) | 0.69 (0.03) | 0.69 (0.02) |

Table 4.6: Performance of logistic regression model using HRV and actigraphy features for eight-week outcome prediction. Results are reported as mean (standard deviation).

| **Outcome** | **Acc.** | **AUC** | **TPR** | **TNR** | **PPV** |
|---|---|---|---|---|---|
| PTSD | 0.56 (0.03) | 0.56 (0.05) | 0.58 (0.06) | 0.53 (0.06) | 0.55 (0.02) |
| PTSD, Sleep Anx./Panic | 0.58 (0.05) | 0.61 (0.06) | 0.64 (0.07) | 0.53 (0.08) | 0.58 (0.08) |
| PTSD, Pain Int. | 0.65 (0.04) | 0.70 (0.02) | 0.69 (0.04) | 0.63 (0.08) | 0.65 (0.08) |

*Results of Experiment 3*

For comparison with the fusion models, experiments were repeated on the participants who contributed both research watch and survey data. The AUC was improved for participants whose PTSD status has changed, in all outcome types compared to the ED survey only models. For PTSD outcome AUC improvement was two percentage points. For PTSD-Sleep Panic/Anxiety outcome, improvement was six percentage points, and for PTSD-Pain Int. outcome improvement was 26 percentage points. The AUC of the overall model (including all participants) was also improved to 0.79 for PTSD-Pain Int. outcome type as shown in Table 4.7. However, for all outcome types, AUC of survey and fusion models were not significantly different as determined by Hanley and McNeil two-tailed test.

Table 4.7: AUC comparison of different model types. Results are reported as mean (standard deviation).

| Model | PTSD | PTSD-Sleep Anx./Panic | PTSD-Pain Int. |
|---|---|---|---|
| Survey | 0.74 (0.03) | 0.77 (0.07) | 0.75 (0.09) |
| Research watch | 0.54 (0.04) | 0.55 (0.08)* | 0.68 (0.04) |
| Fusion | 0.73 (0.04) | 0.75 (0.09) | 0.79 (0.04) |

* Table entry corrected to show the result for PTSD-Sleep outcome.

### 4.2 Change-point decoder sleep detection performance on the AURORA dataset

CPD method described in section 3.2 was also applied to AURORA dataset. One of the challenges in applying sleep detection methods to wearable data collected in-the-wild is accurate detection of the sleep period. If the sleep period is detected accurately, it could be used to estimate metrics related to sleep. For example, sleep efficiency is the ratio of total sleep time to the sleep period and it requires knowledge of both the sleep period length and the sleep/wake estimate during the sleep period. Asking the participants to fill sleep diaries is one method of recording sleep periods in research studies [115]. However, this technique could be subjective as it requires the participant to input sleep onset and offset times and this procedure was not was not used in AURORA study. In addition, Verily research watch did not record time zone information. Therefore, two different methods to estimate the sleep period were tested. For the first method, the rest periods obtained by fitting a cosine (described in detail in subsubsection 4.1.2) was used as the sleep periods. For the second method, the time zones of participant's ED visit during the enrollment was recorded. It was assumed that the participant stays in this timezone during the entire study period. After the detection of sleep periods, the data within each period was used for sleep detection and sleep feature estimation.

The data from each sleep period was converted to change-points. Then, CPD method was applied to change event time series. As described in subsection 3.2.6, the model was not retrained and the weights obtained from the Emory PTSD dataset was used directly. Two different features were calculated to quantify the sleep disturbance. Sleep efficiency ($SE$) was used to quantify how well participant slept. The number of sleep/wake transitions ($SWNum$) was calculated to assess the frequency of episodes of wakefulness and sleep disturbance.

To assess the severity of sleep disturbance, a slightly modified version of the Insomnia Severity Index was administered at week eight of the study [116]. Lower scores indicate no clinically significant insomnia while higher scores indicate severe clinical insomnia. Participants who had contributed at least more than three sleep periods were determined and by using this questionnaire, the participants with severe insomnia were assigned to the first class (N=82) while the remaining participants were assigned to the second class (N=411). Majority class was randomly under-sampled to obtain balanced classes. The mean, standard deviation, and the slope of sleep features (*SE*, *SWNum*) were calculated from the 14-day window before the Insomnia Severity Index questionnaires. A logistic regression model with L1 regularization was built to estimate the outcome classes determined by the questionnaire. Five-fold cross validation was used to validate the model.

Table 4.8: Performance of sleep outcome prediction models. Results are reported as mean(standard deviation) of the external folds of each experiment.

| **Sleep period estimation technique** | **Acc.** | **AUC** | **TPR** | **TNR** | **PPV** |
|---|---|---|---|---|---|
| Cosinor-based | 0.54 (0.05) | 0.58 (0.06) | 0.57 (0.13) | 0.51 (0.09) | 0.52 (0.04) |
| ED time zone | 0.59 (0.06) | 0.61 (0.08) | 0.60 (0.09) | 0.58 (0.16) | 0.58 (0.08) |

Table 4.8 shows the performance of the models for week eight outcome prediction when different sleep period estimation techniques are used. The ED time zone based sleep period model achieved 11% better performance compared to the baseline and these results indicate that this technique holds promise to quantify the sleep disturbance of the participants. It can also be seen that the performance depends on the selection of sleep period estimation technique.

## 4.3 Unsupervised feature extraction using convolutional variational autoencoders

Deep learning methods could learn meaningful representations from the actigraphy data. Furthermore, unsupervised learning is well suited for this problem space because the data without the labels derived from clinical surveys could be utilized. This research represents the first attempt to apply unsupervised deep learning methods to actigraphy data for feature extraction and representation.

The outcomes of three clinical surveys were used to create the classes for the binary classification experiments. The Post-traumatic Stress Disorder Checklist for DSM-5 (PCL-5) was used to measure PTSD symptoms and participants with PCL-5 score greater than 28 were labeled as PTSD [117]. Scored depression variables from PROMIS Depression - Short Form 8b (PROM-Dep8b) were used to measure depression symptoms, with a threshold of 60 for depression [118]. One item from Pittsburgh Sleep Quality Index Addendum [108] ("how often did you awaken from sleep with severe anxiety or panic") was used to assess the difficulty in staying asleep (PanicSleep). In the first experiment, the PCL-5 survey was used by itself, while in the second experiment, PCL-5 and PanicSleep survey scores were combined to determine the outcome classes. Lastly, all three surveys were combined to find the participants who experienced both depression and PTSD symptoms. Also, the participant's general physical health status in the 30 days pre-trauma was used as a feature to test if it could increase the performance of models based on passive actigraphy data. This pre-health score is a derived normative score based on questions from the 12-Item Short Form Health Survey (SF-12) [119].

In the experiments, all participants from the minority class were used, and the over-represented majority class was under-sampled so that the results were not biased due to the unequal class prevalence. For the deep learning models, internal 5-fold cross-validation was performed. All experiments were repeated 30 times on external folds with different random samples from the majority (healthy) class.

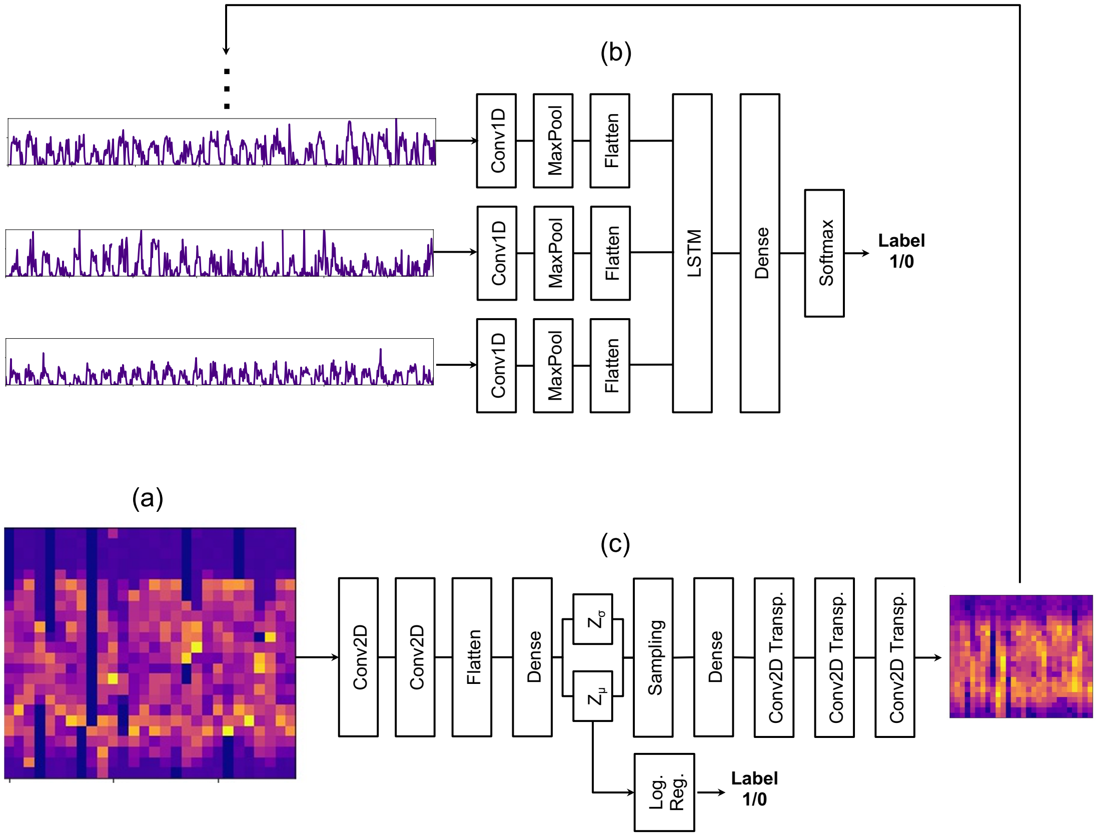


Figure 4.11: (a) Daily actigraphy levels for one participant, this double plot shows activity levels measured over 28 days. (b) CNN-LSTM model, movement data with 24-hour time steps are used as inputs to time distributed 1-D CNN layers. (c) VAE model, latent features were used with a logistic regression classifier for outcome prediction. VAE was also used to generate artificial movement data to train the CNN-LSTM model.

A variational autoencoder (VAE) is an unsupervised generative model that has an encoding phase in which the input data is projected onto lower-dimensional latent representations and a decoding phase that reconstructs the input, as shown in Figure 4.11 (c). However, in the VAE model, the encoder is trained under the restriction that the latent representations follow a Gaussian distribution $N(Z_{\mu}, Z_{\sigma})$. In this work, unlabelled movement maps were used to train a convolutional VAE with two 2D convolution layers (Conv2D) with 16 and 32 number of filters and kernel sizes of 3. The number of units in the dense layers was set to 16. The number of filters in the Conv2DTranspose layers were 32, 16, and 1. The embedding dimension of VAE was 8. The model was trained for 30 epochs with a batch size of 128. Then, the latent representation of the movement maps ($z_{act}$) was used as input features to a logistic regression model in binary classification experiments.

Secondly, an alternative supervised CNN-LSTM model was trained to estimate mental health outcomes from clinical surveys. The number of filters in the Conv1D layer was set to 32, and the kernel size was 3. The number of units in the LSTM and the dense layer was set to 20. Actigraphy data were inputted as 24-hour subsequences, and the model was trained for 30 epochs with a batch size of 32. Lastly, 100 healthy and 100 unhealthy artificial movement maps were generated with VAE models by using randomly sampled encoding vectors. The artificial data was used in the training step of the CNN-LSTM model to test if the performance will be improved.

We visualized what the different VAE latent representations learned by plotting their traversals as shown in Figure 4.12.

We observed that when the unsupervised features extracted with the VAE model were combined with the physical health before the traumatic event (captured with the $SF-12$), the model achieved an AUC of 0.64 and an accuracy of 0.60 in differentiating healthy participants from participants showing PTSD and depression symptoms as determined by clinical surveys. When model was reduced to passive data only (by removing the $SF-12$ feature), the AUC dropped by 3%, but the accuracy was unchanged. The model performance was also tested

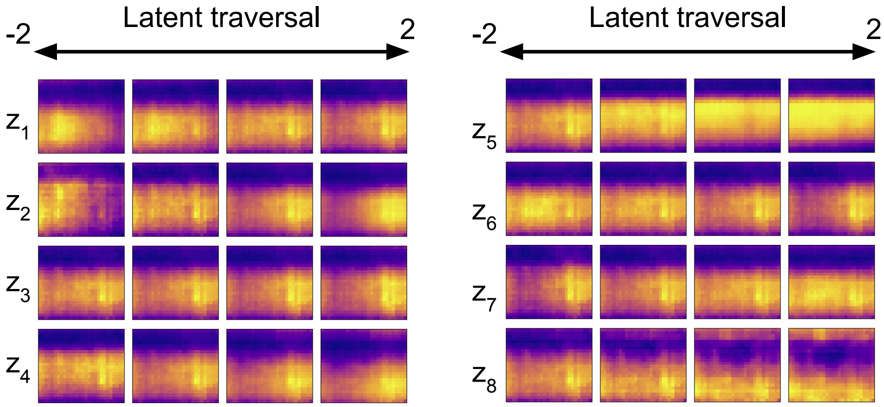


Figure 4.12: Latent traversals of pre-trained VAE model. Each row reconstructs an movement map as the value of each latent dimension is traversed between [-2, 2] while keeping the values of all other latent variables fixed. Most of the variables show decreasing of daily energy, while $z_8$ shows the circadian phase change.

to identify participants with PTSD and sleep disturbance, as shown in Table 4.9. The CNN-LSTM model had an accuracy of 0.56 and an AUC of 0.57 in classifying healthy participants and participants showing PTSD and depression symptoms. By incorporating the artificial data generated by the VAE, the recall of the model increased from 0.45 to 0.60, while other metrics did not change.

Table 4.9: VAE model mental health outcome estimation performance. Results are reported as mean(standard deviation) of the external folds of each experiment.

| **Outcome** | **Features** | **Acc.** | **AUC** | **Precision** | **Recall** |
|---|---|---|---|---|---|
| PCL-5 | $z_{act}$ | 0.55(0.01) | 0.56(0.01) | 0.55(0.01) | 0.51(0.01) |
| PanicSleep<br>PCL-5 | $z_{act}$ | 0.59(0.02) | 0.61(0.03) | 0.59(0.02) | 0.57(0.02) |
| PanicSleep<br>PCL-5<br>PROM-Dep8b | $z_{act}$ | 0.60(0.03) | 0.61(0.03) | 0.60(0.04) | 0.58(0.03) |
| PanicSleep<br>PCL-5<br>PROM-Dep8b | $z_{act}$<br>SF-12 | 0.60(0.02) | 0.64(0.03) | 0.61(0.02) | 0.57(0.03) |

Table 4.10: CNN-LSTM model mental health outcome estimation performance. Results are reported as mean(standard deviation) of the external folds of each experiment.

| Outcome | Features | Acc. | AUC | Precision | Recall |
|---|---|---|---|---|---|
| PCL-5 | $z_{act}$ | 0.52(0.01) | 0.53(0.01) | 0.53(0.02) | 0.40(0.04) |
| PanicSleep<br>PCL-5 | $z_{act}$ | 0.56(0.03) | 0.59(0.03) | 0.58(0.04) | 0.48(0.06) |
| PanicSleep<br>PCL-5<br>PROM-Dep8b | $z_{act}$ | 0.56(0.04) | 0.57(0.04) | 0.58(0.05) | 0.45(0.08) |

## 4.4 Comparison of models using movement data

In order to directly compare all approaches for the movement data (feature extraction, VAE-based feature extraction, and CNN-LSTM), the movement features from subsection 4.1.3 were used to estimate the same outcomes as the deep learning methods. This movement data based feature set included $IS_{rest}$, $IS_{act}$, $IV_{rest}$, $IV_{act}$, $MV_{rest,\mu}$, $MV_{rest,\sigma}$, $MV_{act,\mu}$, $MV_{act,\sigma}$, $CRS_{\mu}$, $CRS_{\sigma}$, $RSI_{\mu}$, $RSI_{\sigma}$, $Mesor_{\mu}$, $Mesor_{\sigma}$, $Amplitude_{\mu}$, $Amplitude_{\sigma}$, $Phase_{\mu}$, $Phase_{\sigma}$ and more details about each feature could be found in Table 4.1. For PTSD outcome, model performed slightly above random chance (Acc.=0.53 and AUC=0.55) similar to previous analysis. For PTSD-Sleep Panic/Anx. outcome, the model achieved an accuracy of 0.60 and an AUC of 0.62. For PTSD-Pain Int., accuracy of the model was 0.59 and the AUC was 0.64. On the other hand, when the VAE-logistic regression approach was run to estimate the PTSD-Pain outcome the accuracy was 0.52, AUC was 0.53. In comparison, CNN-LSTM model's accuracy was 0.52, and AUC was 0.51 for this outcome.

## 4.5 Discussion

In this part of the dissertation, features and patterns related to ultradian and circadian rhythmicity derived from data recorded on a research watch were used to predict or detect post-trauma outcomes. Patients with PTSD have previously reported sleep disturbance symptoms including insomnia and nightmares [49]. Previous studies have also shown that PTSD has a high co-occurrence with chronic pain, which could interfere with patients' daily lives [31, 32]. Moreover, PTSD could also result in decreased interest in activities, as stated by the DSM-5 criteria [30]. Therefore, we hypothesized that PTSD might lead to changes in the biological rhythms that could be captured by the actigraphy and HRV data.

In the first project, three clinical surveys administered in the ED were used as features to train a logistic regression model to predict eight-week PTSD. Using these ED surveys, the models achieved AUCs of 0.73 for PTSD outcome, 0.79 for PTSD-Panic Sleep/Anx. outcome, and 0.77 for PTSD-Pain. Int. outcome. These results indicate that previous PTSD status and stress experienced immediately following the traumatic event are a significant predictor of PTSD in the following months. However, in general, these models predicted that the PTSD status is unlikely to change.

Then, the use of various types of machine learning models with HRV and actigraphy features was investigated. The logistic regression model achieved the highest cross-validated accuracy for predicting the PTSD label at week eight post-trauma when the data from the enrollment until the end of week eight was considered. The weights of the logistic regression model were analyzed to identify the contribution of each feature Figure 4.9. $NNiqr_{\sigma}$, $avgSQI_{\mu}$, $LF_{\mu}$ and $LFHF_{\mu}$ had the highest relative importance amongst the HRV features. $LF$ power, in particular, was lower in the population with eight-week PTSD (a mean of $1178ms^2$ vs

$1562ms^2$). Since the $LF$ power is dominantly associated with baroreflex activity, it can be interpreted as blunted baroreflex activity over this period [120], which is consistent with the literature on PTSD [121]. Previous studies have also shown that $LF$ power is significantly different in stressful conditions compared to the resting conditions [122, 123]. Therefore, this metric could be reflecting the stress the participants are experiencing following the traumatic event. From the actigraphy based metrics, the movement during the rest and the active parts of the day, $IV_{act}$, $IS_{rest}$, $IS_{act}$ and $CRS_{\sigma}$, metrics were the most important. $IV$ measures the fragmentation of rest/activity rhythm and the transitions between rest and activity, and $IV_{act}$ shows irregular activity during the daytime. $IS$ is a measure of variability between days [124]. $IS_{rest}$ and $IS_{act}$ were informative when the outcome is PTSD-Sleep Anx./Panic. This result could indicate that anxiety resulting from trauma could lead to decoupling from zeitgebers in both rest and activity regions.

It is debatable whether collecting data from surveys or a wearable (such as our research watch) represents a lower burden for subjects who develop PTSD. Wearable technologies such as smartwatches (and even mobile phones) are now commonplace and provide the opportunity to collect data without user intervention, while survey-based assessments are active data collection techniques requiring effort and input from the user. However, wearables also require frequent device charging at regular intervals, which is unsustainable in the long term unless a user already is in the habit of doing so. It may not be an either/or proposition, though, and these two approaches could complement each other. For example, participants who were not able or willing to fill in the survey at admission could benefit from passive data collection. In our study, $N = 533$ participants did not fill the ED surveys, but they wore research watches. For these participants, watch-based models could become the prime monitoring method. However, compliance could also be affected by diagnostic status. Research watch data compliance was calculated as the hours with data divided by total hours in the eight-week window, and it was significantly different in PTSD-Sleep Panic/Anx. groups as determined by the Wilcoxon rank sum test. Average compliance was 83% for the first group (PanicSleep $\geq 3$ and $PCL - 5_{week8} \geq 31$) and was 86% for the second group (PSQIA-PanicSleep $< 3$ and $PCL - 5_{week8} < 31$). The compliance to ED surveys (PDI, MCEPS, $PCL - 5_{ED}$) was higher for $PCL - 5_{week8} \geq 31$ group (69%) compared to $PCL - 5_{week8} < 31$ group (48%), and this difference was statistically significant as determined by the Fisher exact test. The research watch models could be more useful for participants undergoing a change in clinical status since the data analysis is windowed and can provide a daily or weekly output which may be interpreted as the severity of illness. This could facilitate the evaluation of response to intervention, for example. Therefore, watch-based models have the potential for passive monitoring over long study periods.

The cosinor method described in this work for determining the rest and activity regions could be helpful for the studies in which participants across different time zones or in situations when obtaining sleep diary and time zone information would be highly burdensome for the participant. In the second sub-project, these rest regions were used as the sleep periods, and sleep-related metrics were derived from the data within the periods by applying CPD sleep detection technique. However, the model achieved only slightly better performance than random chance (AUC=0.58) to detect severe sleep disturbance. In comparison, time zones recorded during the ED enrollment were assumed to be the time zone participants were located, and then CPD was applied to the 6-hour window after midnight. This second technique performed better (AUC=0.61) compared to cosinor-based rest period detection, indicating the importance of sleep period estimation method selection. However, in both cases, the performance of estimating the self-reported insomnia outcome was low. This finding is consistent with previous research [5, 125], and could indicate that self-reported insomnia does not match the objective estimates for the PTSD patients.

In the third sub-project, unsupervised learning was leveraged, and it was shown that the VAE model could extract more informative features and achieved higher accuracy compared to the CNN-LSTM. VAE approach compresses four-week worth of actigraphy data into an 8 dimensional vector. Therefore, this method could result in immense memory savings for applications with more data streams or long-term studies and could be adapted to different and novel devices. Finally, the performance of the deep learning models and machine learning models with hand-extracted features were compared using the motion data. By looking at the results, it can be seen that estimating the PTSD outcome based on only PCL-5 survey is difficult for all models. The performance was better when somatic symptoms such as sleep or pain are included, and hand-crafted features perform slightly better compared to deep learning-based approaches. As more data is collected, the training of deep learning methods could be improved.

There are several limitations to this work. First, the outcomes (PTSD status at week eight) may reflect the appearance of PTSD at any time over the intervening eight weeks. The high variability in the speed of development of PTSD is likely to create high class confusion in any machine learning paradigm. Moreover, there is the potential for individuals' PTSD symptoms to wax and wane over eight weeks, further confusing any algorithm trained on such data. Second, due to the use of self-report surveys from week eight for constructing outcome classes, our cohort is a subset of the original AURORA Freeze 2 dataset, albeit a rather large cohort. As more data are collected in the AURORA study in the coming years, we will address this limitation by re-evaluating the methods with more participants. Lastly, time zone information was not available for our participants. Circadian (mis)alignment may have provided additional information for adjusting features. While the cosinor-based rest-activity detection might compensate for this lack of information, it cannot fully address the issue.

## 4.6 Conclusion

In this chapter, various approaches to quantify ultradian and circadian rhythms are presented. This research represents, to best of the author's knowledge, the first attempt to predict outcomes following a traumatic event from a wearable (or, more specifically, a research watch). Outcomes were both classified and predicted using non-invasive physiological features derived from the research watch, using a logistic regression model. A method to automatically detect rest and activity periods of the day was also developed using the cosinor analysis and combined with CPD method to estimate sleep metrics. Activity heat plots were used for unsupervised feature extraction for the first time using deep learning methods.

Acute conditions could lead to chronic conditions without the required treatment and they also have the capacity to improve with early detection of symptoms. For example, accurate prediction of PTSD in the early aftermath of trauma would enable early preventive interventions [126]. Rothbaum *et al.* [127] showed that trauma survivors receiving an early modified prolonged exposure intervention reported significantly less PTSD severity compared to the assessment group. It has also been shown in a preliminary study that administering an early single high-dose hydrocortisone could reduce the risk of PTSD development [35]. While wearable-based PTSD detection based on only PCL-5 survey was not very effective, the performance was very promising when other somatic symptoms such as sleep disturbance or pain was included in the outcome classes. Notably, the model for participants with a combined PTSD and pain outcome combined provides the highest performance. Identifying and treating these particular types of individuals is extremely important. Previous studies report that patients with both chronic pain and PTSD combined use healthcare services more than the patients with PTSD or chronic pain alone, increasing healthcare costs [32]. Moreover, PTSD treatment for these patients could be more beneficial than for other groups, since they also report a reduction in pain symptoms after treatment [128].

In conclusion, the methods for classifying or predicting outcomes (for window sizes smaller than 49 days) could be useful in passively monitoring changes in symptom severity in large populations and in low-resource settings. Without the prior knowledge of which patients to administer treatment to, smartwatch-based monitoring could be used to identify the subset of patients to prioritize.

# CHAPTER 5
# CHRONIC DISEASE STATE PREDICTION

Chronic diseases are health conditions that are persistent and last long periods of time with reoccurring symptoms [129]. In this part of the thesis, the signal processing and machine learning techniques for chronic disease state prediction is presented. Using the dataset described in subsection 2.6.3, features were derived from passive and active data collected by the smartphone-based framework for predicting or classifying heart failure decompensation events. The movement data from research watches in DMD and Emory PTSD studies were also analyzed in relation to clinical outcomes. Figure 5.1 shows the datasets used in this part of thesis on the disease time scale plot.

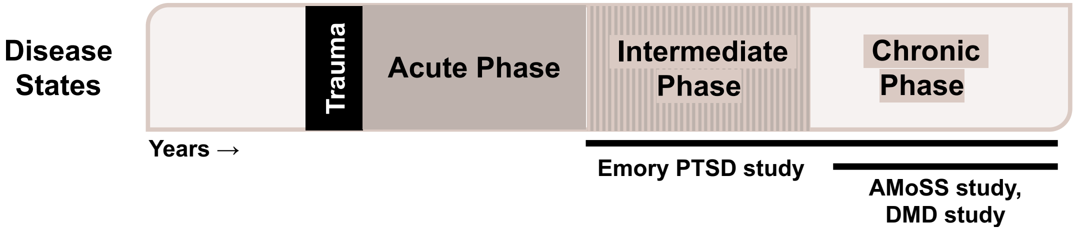


Figure 5.1: Illustration of AMoSS, DMD and Emory PTSD studies on disease time scales.

## 5.1 Methods for monitoring heart failure patients using passive smartphone data

### 5.1.1 Data collection using the smartphone application and the clinical events

Figure 5.2 illustrates the study timeline and the data collection after the enrollment and discharge from the clinic. We analyze different data modalities, including motion, social contact, location, and clinical survey data (KCCQ) collected by the smartphone. We develop algorithms based on using a single modality and two different sensor fusion approaches. We also present an analysis of the feature importance in the model and report a novel late-fusion model which combines the KCCQ, motion, and social contact data.

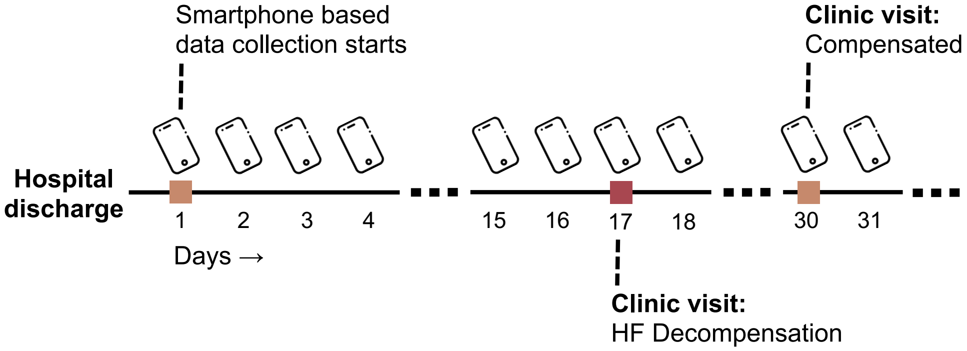


Figure 5.2: Illustration of the AMoSS HF study timeline. Passive data collection started after the hospital discharge, and the clinical team recorded the clinical events after the enrollment.

*Clinical events*

Clinical events consisted of decompensated and compensated events and were collected by the clinical team when the participants visited the hospitals. In the compensated events, the participants visited the hospital for any reason, and their fluid levels were determined to be normal. For the decompensated events, the clinical team determined the participant to have functional limitation related to HF. Decompensated and compensated events were assigned to positive and negative classes respectively. The number of events contributed by each participant varied as shown in Figure 5.3.

*Passive data sources*

The raw 3D accelerometer data was converted to activity counts using the Actigraphy Toolbox to reduce the required memory for storing and eliminate noise [130].

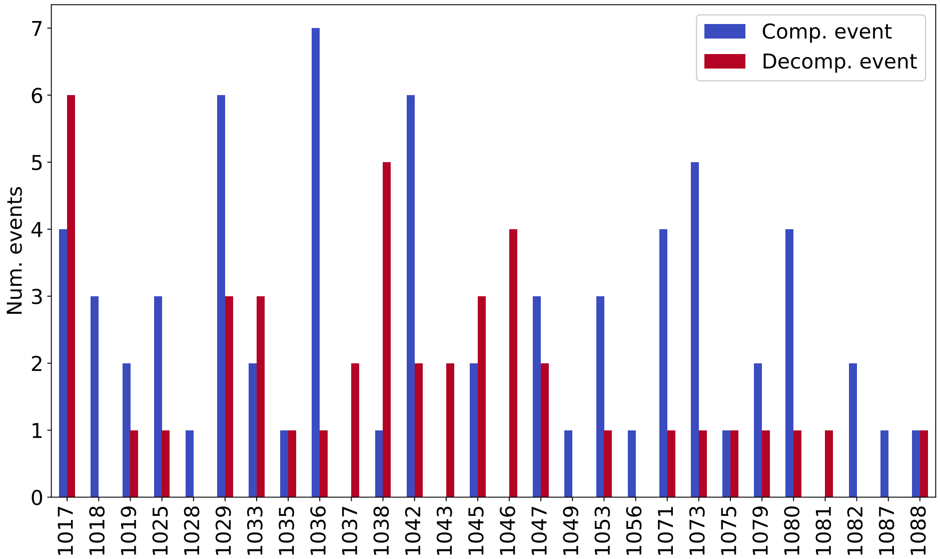


Figure 5.3: Number of events for each participant. y axis shows the number of events and x axis shows the unique IDs. Compensated and decompensated events are shown with different colors.

In the first step, the z-axis of the accelerometer data was filtered using a band-pass Butterworth filter with $0.25 - 11$ Hz passband to eliminate extremely slow or fast movements [131]. Then, the maximum values inside 1-second windows were summed for each 30-second epoch to obtain the activity counts, following the approach described by Borazio *et al.* [98]. Figure 5.4 illustrates the double plot for the motion data for one participant over a recording period of 300 days. White regions indicate missing data, which could be due to the participant turning off the data sharing or the smartphone running out of battery.

Social contact data included the call data and the duration of each call. Each contact was anonymized and assigned a unique identifier at the source. Figure 5.5 illustrates one participant's social contact over 300 days for the ten most frequently contacted IDs. Lastly, location data was collected using the Android location services application program interface, which generally used cellphone tower or WiFi and not GPS for geolocation. Figure 5.6 shows the location data of

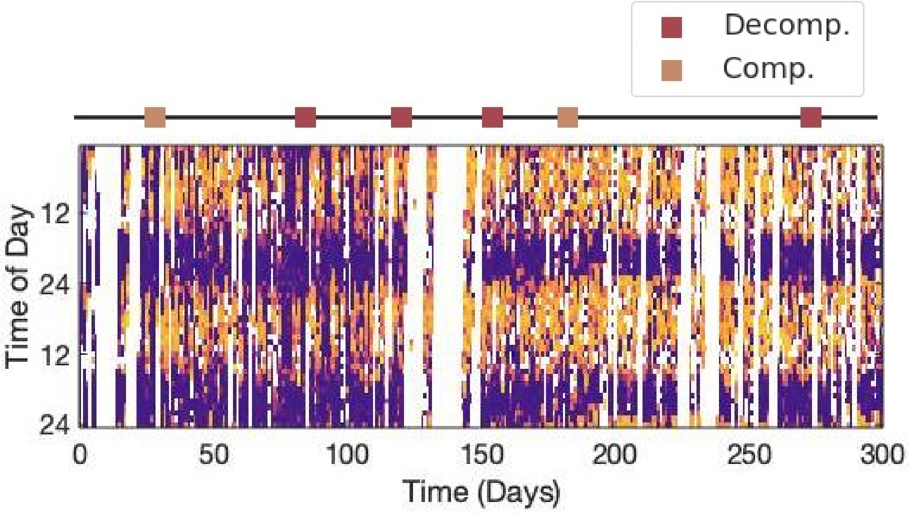


Figure 5.4: Double plot representation of smartphone actigraphy data, which illustrates daily motion intensity levels for one participant. Darker colors indicate lower intensity movement, and the white color indicates missing data. On the top of the plot, decompensated and compensated clinical events are shown with red and orange squares respectively.

a participant, collected from compensated and decompensated windows. High spatial resolution was not required since the aim was to identify the general environment in which a user was located. (E.g., home, work, shops, etc.) If the smartphone moved at least 100 meters, and at least 5 minutes had passed since the last location data update, a new relative location was recorded. These parameters were defined while designing the app to preserve battery life while still providing sufficient temporal and spatial resolution in comparison to the phone's ability to geo-locate without GPS.

*Active data sources*

Active data type, which required user input, was KCCQ administrated through the smartphone app. The scores are lower for severe HF symptoms, and KCCQ scores $\leq 25$ correspond to New York Heart Association (NYHA) class IV. In this study, we used the shorter version of the questionnaire, referred to as KCCQ-12 [132]. The KCCQ-12 survey had physical limitation, symptom frequency, qual-

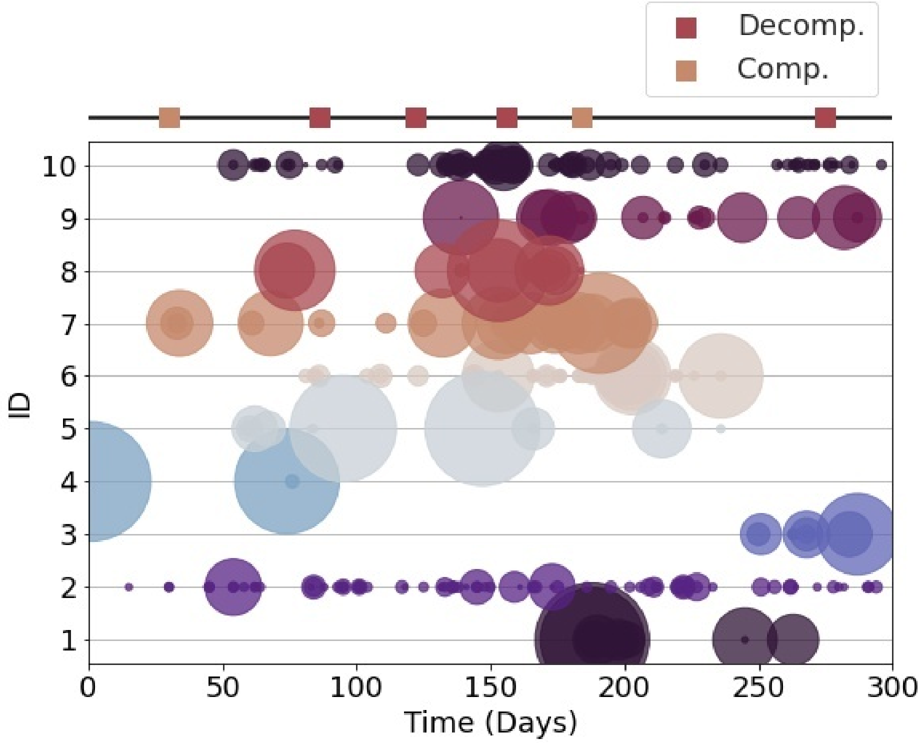


Figure 5.5: Participant's social contact intensity over 300 days. Each unique contact is assigned a number as shown in the y-axis, and the circle radius is proportional to call duration to each ID. On the top of the plot, decompensated and compensated clinical events are shown with red and orange squares respectively.

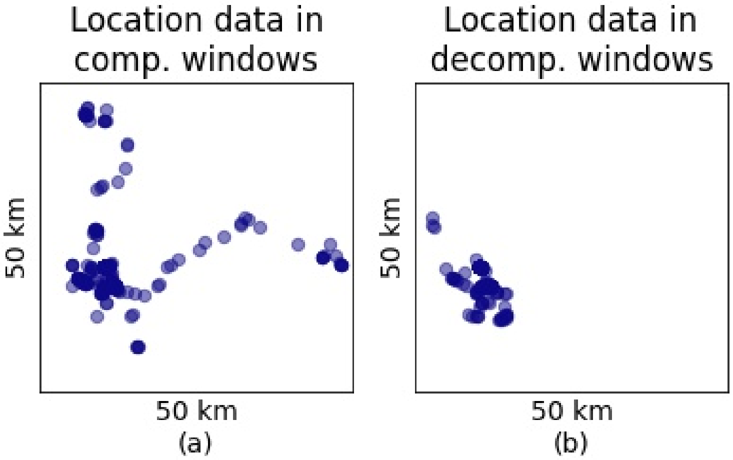


Figure 5.6: Location data collected in compensated and decompensated windows, shown on the same map with 50x50 km dimensions.

ity of life, and social limitation domains, and the summary score (ranging from 0-100) was the average of all domains. Figure 5.7 shows the KCCQ-12 scores administrated through the app.

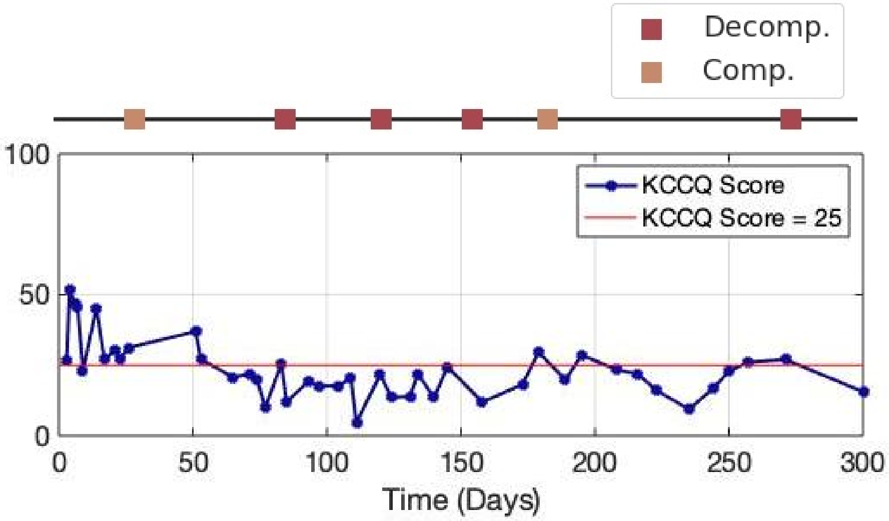


Figure 5.7: KCCQ summary score over days for one participant. KCCQ score $\leq 25$ indicates a transition to severe HF. On the top of the plot, decompensated and compensated clinical events are shown with red and orange squares respectively.

### 5.1.2 Personalized prediction models of heart failure severity

In the preliminary analysis conducted in 2018, the feasibility of predicting the quality of life of patients with HF using passive smartphone data measured via the smartphone app was assessed [64]. Patient-specific models to estimate KCCQ using various features from past passively monitored data were built. In a subsequent analysis, these features were used to cluster patients into high (KCCQ $\leq$ 25) versus moderate (KCCQ $>$ 25) severity HF groups.

*Feature extraction*

Several features were extracted from movement data to evaluate rest-activity characteristics and circadian rhythms of subjects. Rest-activity rhythms were assessed using IS, IV, Most Active 10 Hours (M10), and Least Active 5 Hours (L5) [124].

IS and IV were described in detail in subsection 4.1.3. M10 is the average of the most active 10 hours over all days. A drop in M10 could imply a reduction in physical activity due to HF symptoms. Lastly, L5 is the average of the least active 5 hours. L5 indicates movements during sleep, and night-time arousal. Cosinor rhythmometry features, total energy of activity data, and the time of day of the maximum activity rhythm were also calculated. The correlation coefficient $R_k$ quantifies the correlation of a variable with itself at a previous time, e.g. $A_i$ versus $A_{i-k}$. In this work a lag $k = 24$ hours was used. A more pronounced circadian rhythm will result in a higher $R_k$ [133]. MSE was calculated to quantify irregularity or unpredictability of behavior over multiple timescales. MSE was calculated following the methods described by Costa *et al.* [134]. Actigraphy time series were coarse-grained by averaging the data points within non-overlapping windows. The first 20 scales of multiscale entropy were calculated by varying the window size from 1 to 20. For each coarse-grained time series, sample entropy was calculated.

Severe HF causes discomfort and can hinder physical activity, which could lead to the subject staying at home more, or altering routine behaviors. Location features were extracted to capture these changes. Using all location data from each subject, the "home" location was defined as the most frequently visited location. The percentage of time spent at "home" was calculated. The second most frequently visited location was determined in a similar fashion as "home". The percentage of time spent at the second most frequent location was calculated.The area within a 20 km radius from home was designated as "zone-1", and the area outside of this radius was called "zone-2", as shown in Figure 5.8. The number of times the subject visited each zone was counted.Haversine distances between all locations to the "home" location were summed. The Haversine distance is the shortest distance between two coordinates over the surface of the earth.

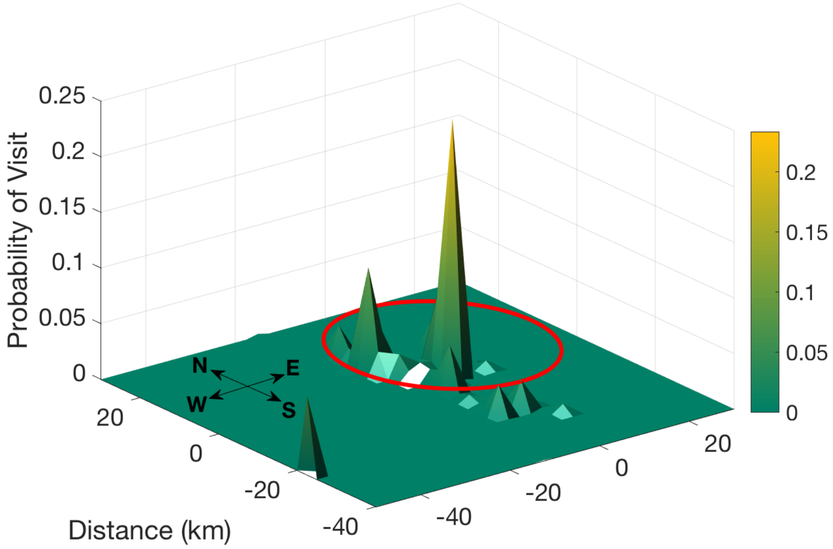


Figure 5.8: Example of location data collected with AMoSS app. Increasing height, represents the probability of visiting an area. Areas south and west from the origin are represented as negative distances. The red circle is the boundary of "zone-1", the area enclosed by a circle of 20 km radius from most frequently visited location.

HF symptoms could affect a subject's social behavior. The following contact activity features were extracted from smartphone data: total number of calls, mean duration of calls, standard deviation (std) of duration of calls, mean duration without any calls, standard deviation of duration without any calls.

*Machine learning models*

Personalized models were created for each subject to estimate the KCCQ summary score. GLM with binomial distribution and logit link was used. Elastic net regularization was applied to personalized GLMs to decrease bias and improve classification performance [135]. Models were built only for subjects from whom sufficient data was gathered over at least ten windows at that time of the study, and a window was defined as two weeks. If insufficient data was gathered from ten or more windows, or did not share a specific data type at all, no model was built for that data type.

Model performance was assessed via record-wise leave-one-out cross validation (LOOCV). Given $N$ windows of data, $N-1$ windows are used to train the model for a given patient for predicting KCCQ scores, and the held-out window is used as the test set data from which a KCCQ score is predicted. This is repeated for the remaining $N-1$ windows. The Mean Absolute Error (MAE) between actual and estimated KCCQ scores was calculated for each personalized model.

In addition to regression analysis, a classification analysis with a K-Nearest Neighbors (K-NN) approach was performed after quantizing KCCQ summary scores. Data were dichotomized into KCCQ scores $\leq 25$ or $> 25$ [136]. KCCQ scores $\leq 25$ correspond to New York Heart Association (NYHA) class IV. Patients with class IV HF are unable to complete any physical activity without discomfort. KCCQ scores $> 25$ correspond to NYHA class I-III which describes less severe HF compared to class IV. Clustering analysis was performed for two subjects who had enough KCCQ summary scores in each class. Cosine distance and five nearest neighbors were used as model specifications. Five-fold cross validation was implemented whereby the model was trained on four folds and the fifth held-out fold was used for testing, and this process was repeated for the remaining four folds. The percentage of correctly classified points were reported for each subject.

### *Experimental results*

The average MAE over the population for estimating the KCCQ using only activity metrics was 5.71 units (or percent, since the scale is normalized to be between 0-100) as shown in Table 5.1. When only location metrics were used to estimate the KCCQ scores, the MAE rose to 7.40 (N=8). For personalized models based on contact activity features, the population average (N=9) MAE was 6.05. For personalized models built with features from all three data domains, the population average MAE was 5.43. Although one might therefore be tempted to infer that movement provides the most information in this context, inspection of Table 5.1 shows that the most useful type of data depends on the individual. Moreover, the most frequently selected features by elastic net in the LOOCV procedure also varied according to the subject. This creates a strong case for personal models trained on initial KCCQ reports. It can also be noted that the models developed here outperform the baseline (sample-and-hold) estimate, i.e. simply using the first KCCQ score. No error exceeded 8% when all three domains of data types were available.

Table 5.1: Mean absolute error of leave-one-out cross validation KCCQ estimation models for each of the first 10 patients in the AMoSS HF dataset.

| **Initial KCCQ score** | **Motion features** | **Location features** | **Social contact features** | **Combined features** | **Baseline model** |
|---|---|---|---|---|---|
| 95 | 0.43 | 0.90 | 0.91 | 0.44 | 2.12 |
| 50 | † | 2.26 | 2.47 | 2.25 | 40.39 |
| 13 | 4.93 | † | 4.01 | 3.92 | 4.74 |
| 56 | 5.45 | 7.73 | 6.44 | 5.45 | 34.86 |
| 71 | 5.46 | 5.44 | 5.68 | 5.49 | 6.61 |
| 38 | 7.50 | 12.08 | 7.41 | 7.52 | 17.96 |
| 77 | 9.41 | 9.55 | 7.82 | 7.85 | 19.87 |
| 19 | † | 12.32 | 8.82 | 7.99 | 6.10 |
| 13 | 8.22 | 8.31 | † | 8.33 | 19.31 |
| 33 | † | † | 10.90 | † | 13.32 |
| **AVG** → | 5.71 | 7.40 | 6.05 | 5.43 | 16.53 |

† indicates patient did not share data type.

Figure 5.9 provides a visualization of points with t-Distributed Stochastic Neighbor Embedding [137] in 3 dimensions using social contact features. Using contact activity features from two subjects (Subject 1 and Subject 8 from Table 5.1), five-fold cross validation of the K-NN classification approach was performed. Out-of-sample classification accuracy for these two subjects were 0.78 and 0.88 respectively. Repeating the same analysis with location features resulted in respective classification accuracies of 0.65 and 0.73.

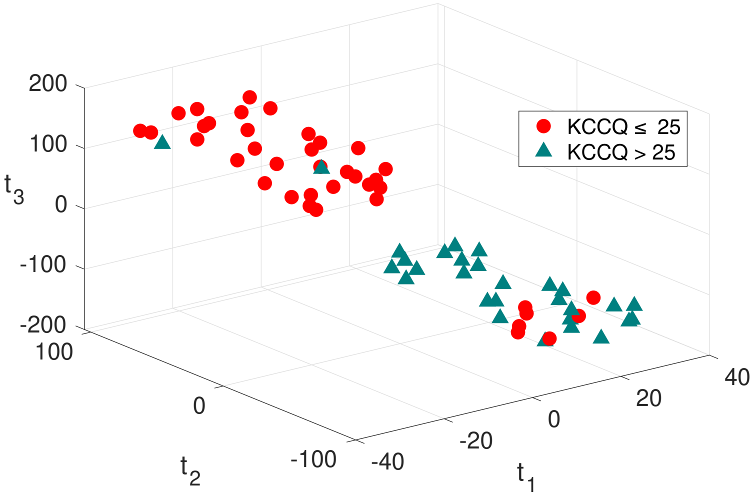


Figure 5.9: Visualization of two KCCQ clusters (Severe HF defined as KCCQ $\leq$ 25) in three arbitrary dimensions (represented by $t_{1-3}$) using t-Distributed Stochastic Neighbor Embedding on social networking behavior features for two subjects.

### 5.1.3 Dynamics of interpersonal social interactions and motion passively captured from smartphones predicts decompensation in heart failure

In the second analysis using AMoSS HF dataset, motion, social, location, and clinical survey data collected via the smartphone-based monitoring system were used to develop and validate an algorithm for predicting or classifying HF decompensation events (hospitalizations or clinic visit) versus clinic monitoring visits in which they were determined to be compensated or stable. Models based on single modality as well as early and late fusion approaches combining patient-reported outcomes and passive smartphone data were evaluated. Passively collected data from smartphones, especially when combined with weekly patient-reported outcomes, may reflect behavioral and physiological changes due to HF and thus could enable prediction of HF decompensation.

*Feature extraction*

Several features were extracted from the data collected through the app to construct the motion feature set. A window of data was the N day period before a clinical event, and the feature extraction was performed for each window. The window size N was chosen to be 14 days initially since it was also selected by the developers of KCCQ to represent the participant's recent functioning [61]. Firstly, from preprocessed smartphone activity counts, descriptive statistics were extracted. These included mean ($act_{mean}$), standard deviation ($act_{std}$), mode ($act_{mode}$), skewness ($act_{skew}$), and kurtosis ($act_{kurt}$). The completeness percentage ($act_{comp}$) was calculated by dividing the epochs with data by the total number of epochs in the window.

For each window, the total number of calls ($numCalls$), the sum of the duration of calls ($durCalls$), the standard deviation of the duration of calls ($durCalls_{std}$), the sum of durations without any calls ($durNoCalls$), and the standard deviation of these durations ($durNoCalls_{std}$) were calculated to be used as social contact features.

Using the participant's location data, the most frequently visited location was determined and defined as the "home" location. The number of times the participant was at the home location was calculated and used as a feature ($atHome$). For the second location feature, Haversine distances between all locations to the home location were summed ($distToHome$). Lastly, the area within a 2 km radius from home was defined as "zone-1". The area outside of this radius was defined as "zone-2". The number of times the participant contributed from these two zones were calculated ($zone_1$, $zone_2$).

From the KCCQ data, two different feature sets were investigated. Firstly, the summation score ($KCCQ_{sum}$) was used as a feature. For the second set ($KCCQ_{all}$), each domain (physical limitation, symptom frequency, quality of life, and social limitation) was used separately. For these two active data feature sets, the performance of using the mean of all surveys inside the window or using the most recent survey was also tested.

### *Machine learning models*

Logistic regression classifiers were trained to map the feature vector to the compensated or decompensated outcome. All the models were written in the Python 3 language, and the programming code was based on Scikit-learn [113]. Since each participant could contribute more than one event, we used leave-one-patient out cross-validation. The model was trained on the data from all participants except one held-out participant, and this participant's data was used as the test set. This process was repeated for each participant in the dataset. All experiments were repeated 50 times to obtain the final results.

Since the number of compensated and decompensated events were highly imbalanced, as seen in Table 2.2, the majority undersampling was performed on the training set before training the classifiers. During the majority undersampling, all participants from the minority class were used, and the same number of participants from the majority class were randomly selected. Sequential forward feature selection was used to select the three most informative features from each modality.

Early and late fusion approaches combined passive and active modalities and are shown in Figure 5.10. In the early fusion approach, extracted features were combined at the input level to create a single feature vector. Secondly, all single modality model's output probabilities were concatenated and used as input to another classifier for the late fusion approach. In the fusion models, the participants who contributed all data types were included in the analysis. Models were also evaluated on the same participants to analyze if the fusion approached improve upon the single modality approaches. Finally, the effect of changing the window size and sliding the window was tested using the same participants for all the models.

Model performance metrics were accuracy (Acc.), Positive predictive value (PPV), True positive rate (TPR), AUC, and Area under the precision-recall curve (AUCPr). In this analysis, positive class was the decompensated events while compensated events were the negative class. Accuracy is the ratio of the number of correctly classified samples to the total number of samples. PPV (or precision) is the ratio of the number of correctly classified positive samples to the number of the samples which is predicted as positive. TPR (or recall) is the ratio of positives that are correctly classified to actually positive samples. AUC is the area under the ROC curve which shows the model performance under different classification thresholds. Lastly, AUCPr is the area under precision-recall curve.

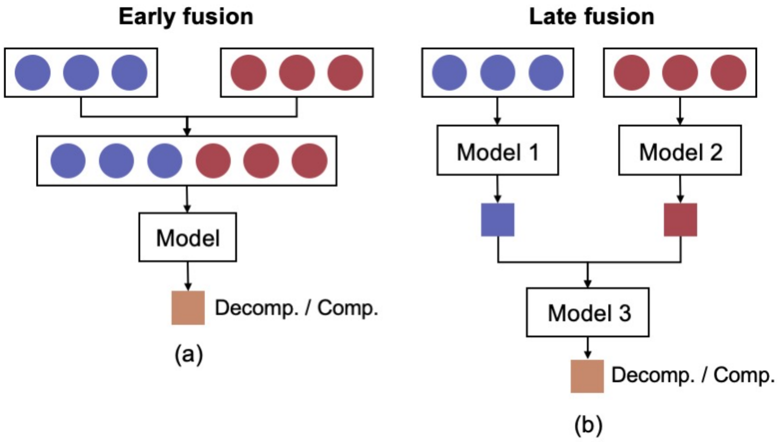


Figure 5.10: Modality fusion techniques. Purple and red colors indicate two different modalities. Figure (a) shows the early fusion approach, and figure (b) shows the late fusion of the modalities.

To examine and interpret the features further, SHapley Additive exPlanation (SHAP) values for the early fusion model were calculated [138]. This framework is model agnostic, and SHAP values quantify the contribution and impact of each feature to the model.

*Experimental results*

The cross-validation performance for each single modality model is shown on Table 5.2. For these experiments, the window was set to 14 days before each clinical event. The number of unique participants and the number of clinical events changed according to the modality since the participants could stop contributing data. For the motion model, 23 participants contributed 28 decompensated events and 44 compensated events. For the social contact model, there were 21 participants with 27 decompensated events and 45 compensated events. Lastly, there were 18 participants with 13 decompensated events and 33 compensated events for the location model.

Table 5.2: Passive data model performances. Results are reported as mean(standard deviation) of the external folds of each experiment.

| **Modality** | **Acc.** | **AUC** | **AUCPr** | **PPV** | **TPR** |
|---|---|---|---|---|---|
| Motion | 0.66 (0.03) | 0.66 (0.03) | **0.60 (0.06)** | 0.55 (0.04) | 0.61 (0.06) |
| Location | 0.59 (0.07) | 0.56 (0.10) | 0.39 (0.11) | 0.34 (0.10) | 0.49 (0.17) |
| Social | **0.58 (0.05)** | 0.65 (0.05) | 0.56 (0.06) | **0.46 (0.06)** | 0.60 (0.07) |

Table 5.3 provides the single modality results for the active data type, KCCQ survey. For two different feature sets ($KCCQ_{sum}$ and $KCCQ_{all}$), the table shows the performance metrics when the mean of all the questionnaires within the 14-day window was used and when the most recent questionnaire was used. For this active data type, 20 unique IDs contributed 23 decompensated events and 32 compensated events. Using the summary KCCQ score and taking the most recent questionnaire has resulted in the highest AUCPr score of 0.69.

In the fusion of KCCQ and motion data, 15 participants contributed 13 decompensated events and 27 compensated events. Similarly, for KCCQ and motion data model, 17 participants contributed data for both modalities, 21 decom-

Table 5.3: Active data single modality model performance. Results are reported as mean(standard deviation) of the external folds of each experiment.

| Modality | Acc. | AUC | AUCPr | PPV | TPR |
|---|---|---|---|---|---|
| Mean of window, $KCCQ_{sum}$ | 0.64 (0.01) | 0.75 (0.01) | 0.61 (0.02) | 0.55 (0.01) | 0.66 (0.03) |
| Mean of window, $KCCQ_{all}$ | 0.65 (0.02) | 0.67 (0.02) | 0.54 (0.04) | 0.57 (0.02) | 0.69 (0.04) |
| Most recent, $KCCQ_{sum}$ | **0.69 (0.01)** | **0.77 (0.01)** | **0.69 (0.02)** | **0.61 (0.02)** | 0.71 (0.03) |
| Most recent, $KCCQ_{all}$ | **0.69 (0.03)** | 0.70 (0.01) | 0.61 (0.04) | 0.60 (0.02) | **0.74 (0.04)** |

pensated events, and 26 compensated events. When three modalities were used (KCCQ, motion, social contact), 16 participants contributed 18 decompensated events and 21 compensated events. Lastly, when all data types were merged, there was data available for 12 participants, ten decompensated events, and 18 compensated events. The results for the early fusion models is shown in Table 5.4 and in Table 5.5 for the late fusion models. The highest AUCPr of 0.77 was achieved when KCCQ and motion and social contact modalities were combined with late fusion. For early fusion models, using the same modalities resulted in an AUCPr of 0.69. The corresponding SHAP summary plot for the early fusion model is shown in Figure 5.11.

Using the best models in each category, how early the algorithm could predict the outcome (time-to-event analysis) was also investigated. In this analysis, participants who shared data for all windows and data types were used (N=13; 13 decompensation events; 18 compensation events). Figure 5.12 illustrates the AUCPr of the models as the window was shifted. From the single modality models, social contact model was most drastically effected by shifting the analysis

Table 5.4: Results of early fusion models. Results are reported as mean(standard deviation) of the external folds of each experiment.

| **Modality** | **Acc.** | **AUC** | **AUCPr** | **PPV** | **TPR** |
|---|---|---|---|---|---|
| Motion, soc. | 0.62 (0.04) | 0.58 (0.03) | 0.54 (0.04) | 0.53 (0.05) | 0.53 (0.06) |
| KCCQ, motion | **0.73 (0.02)** | **0.81 (0.01)** | **0.75 (0.03)** | 0.69 (0.02) | **0.73 (0.05)** |
| KCCQ, motion, soc. | 0.71 (0.04) | 0.72 (0.05) | 0.69 (0.06) | **0.70 (0.04)** | 0.66 (0.09) |
| KCCQ, motion, soc., loc. | 0.67 (0.05) | 0.64 (0.07) | 0.57 (0.11) | 0.55 (0.07) | 0.56 (0.09) |

Table 5.5: Results of late fusion models. Results are reported as mean(standard deviation) of the external folds of each experiment.

| **Modality** | **Acc.** | **AUC** | **AUCPr** | **PPV** | **TPR** |
|---|---|---|---|---|---|
| Motion, soc. | 0.64 (0.03) | 0.63 (0.04) | 0.52 (0.05) | 0.54 (0.04) | 0.56 (0.07) |
| KCCQ, motion | 0.67 (0.03) | 0.75 (0.02) | 0.67 (0.04) | 0.61 (0.03) | **0.72 (0.07)** |
| KCCQ, motion, soc. | **0.71 (0.04)** | **0.79 (0.03)** | **0.77 (0.04)** | **0.68 (0.04)** | 0.70 (0.05) |
| KCCQ, motion, soc., loc. | 0.62 (0.07) | 0.72 (0.07) | 0.60 (0.11) | 0.49 (0.07) | 0.68 (0.10) |

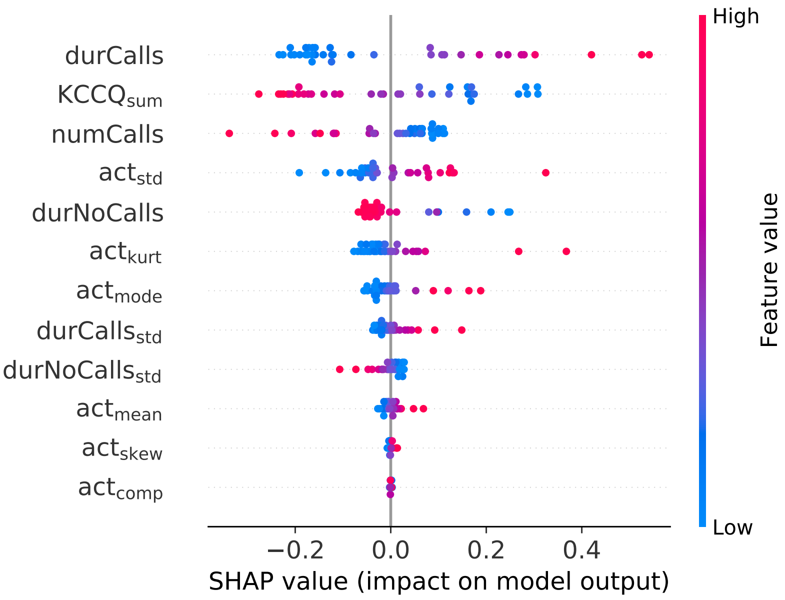


Figure 5.11: SHAP summary plot for the early fusion model. Features are sorted by their impact on the y-axis. Each point on the plot shows the Shapley value for one instance. Horizontal location shows the feature's effect for predicting positive class (decompensated) or negative class (compensated), and color indicates the feature value.

window. Late fusion models performed better compared to KCCQ models when time to event was less than four days. Figure 5.13 shows the effect of the window size to AUC and AUCPr metrics. Similar to time-to-event analysis, only the participants who shared all data types for all windows were used (N=11; 12 decompensation events; 15 compensation events). It can be seen that KCCQ model performance was not effected, since this model used the last score (closest to the event) within the window. Social contact model's performance was improved as the window size was decreased.

## 5.2 Using movement data for chronic post-traumatic stress disorder state estimation

Using the Actiwatch research watch data from the Emory PTSD dataset described in subsection 2.6.1, movement features ($IS$, $IV$, mean and standard deviation of

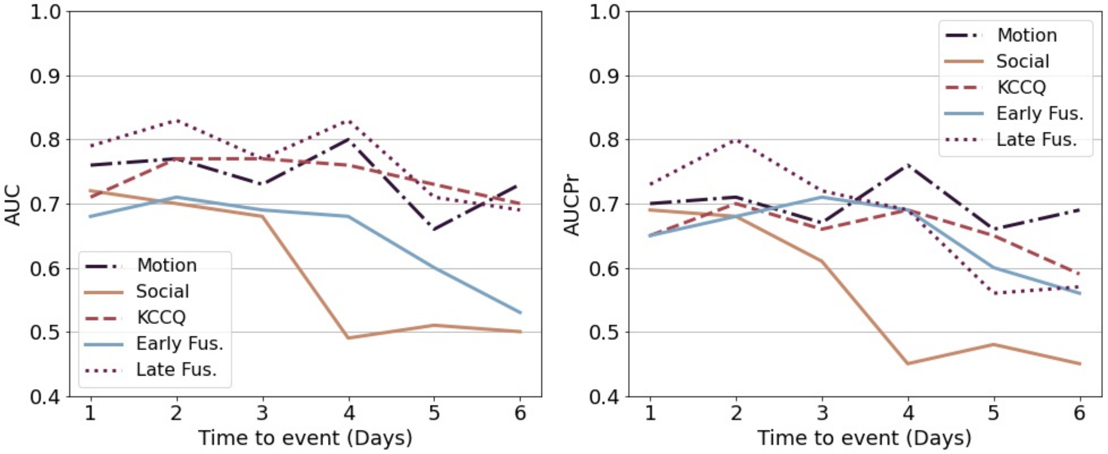


Figure 5.12: Performance change as the data window is shifted. x axis indicates the time-to-event. Early and late fusion models use KCCQ, motion, social contact modalities.

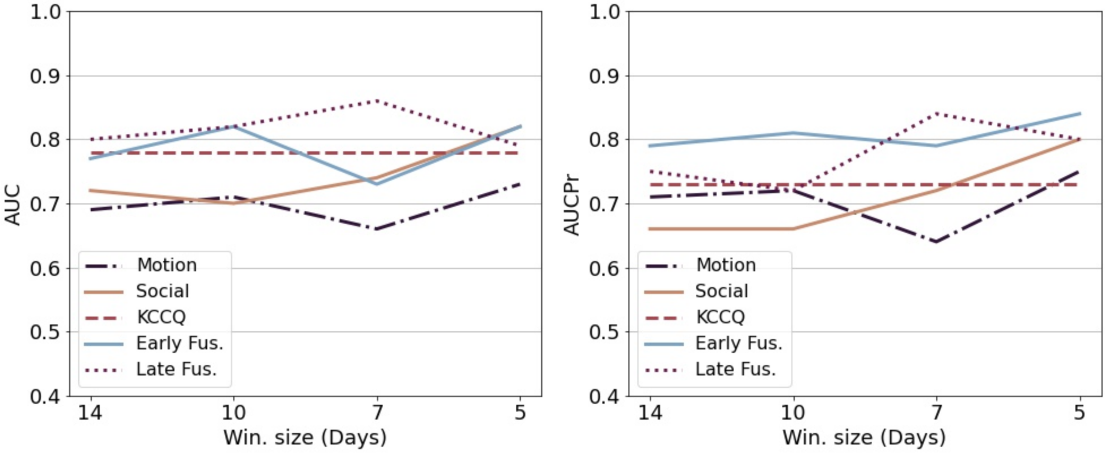


Figure 5.13: Performance change as the data window size is changed. x axis indicates the window size. Early and late fusion models use KCCQ, motion, social contact modalities.

movement, $CRS_{\mu}$, $CRS_{\sigma}$, $RSI_{\mu}$, $RSI_{\sigma}$, $Mesor_{\mu}$, $Mesor_{\sigma}$, $Amplitude_{\mu}$, $Amplitude_{\sigma}$, $Phase_{\mu}$, $Phase_{\sigma}$) were derived. From the participants who wore the Actiwatch, there were $N = 8$ participants who were diagnosed with long-term PTSD and there were $N = 88$. Due to low number of PTSD participants, it was not possible to build a cross-validated classifier. T-test was used to compare the means of features for PTSD and non-PTSD groups. Among these 14 movement features, $CRS_{\sigma}$ and $IS$ features had statistically significant difference ($p = 0.028$ and $p = 0.017$ respectively). Figure 5.14 illustrates the violin plots of $IS$ and $CRS_{\sigma}$ for both groups.

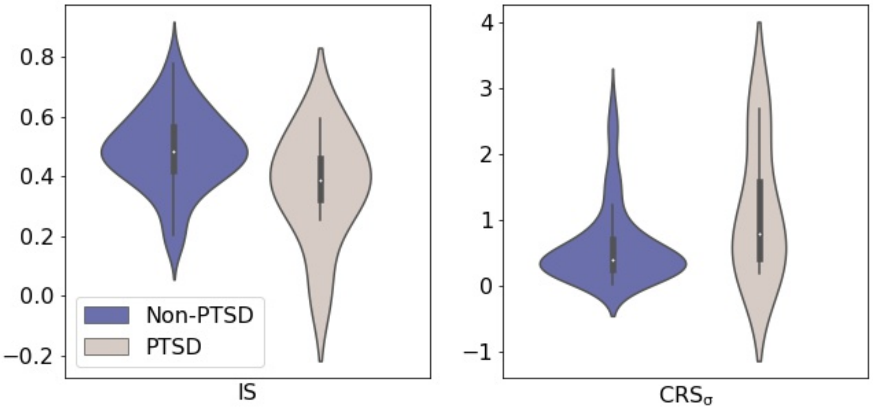


Figure 5.14: Violin plots of motion features $IS$ and $CRS_{\sigma}$ motion features.

## 5.3 Use of a wearable device to assess sleep and motor function in Duchenne muscular dystrophy

In another study, actigraphy data was used to assess the sleep disturbance and day-time activity levels in participants with DMD [139]. A subset of the ambulatory participants (N=13) completed 6-minute walk test (6MWT) during their appointments to assess their functional capacity. The 6MWT measures how far an individual can walk within a 6-minute period and was conducted with ambulatory participants in accordance with the methods described in 2010 by MacDonald *et al.* [140]. The percentage predicted 6-minute walk distance (6MWD) was calculated by dividing the distance walked by the expected distance for each participant's age and height by using the Geiger equation [141]. Parents of all participants completed the Sleep Disturbance Scale for Children [142]. The questionnaire consists of 26 Likert-type questions and one of the factors measured was Disorders of Initiating and Maintaining Sleep (DIMS).

A longitudinal actigraphy dataset was created using the Actiwatch research watch, as described in subsection 2.6.4. The rest/activity metrics (IS, IV, M10, L5) were derived from the research watch actigraphy data of the participants following the steps described in subsection 4.1.3. For the ambulatory group, linear regression models were built to assess the associations of outcomes with the metrics. Within the ambulatory group, linear regression modeling revealed a significant association between M10 and percentage predicted 6MWD such that more daytime activity was associated with better 6MWD performance ($R^2 = 0.41, p = 0.019$). In addition, a more fragmented daily rhythm (higher IV) had more difficulty initiating and maintaining sleep as measured by the DIMS subscale ($R^2 = 0.61, p = 0.008$).

## 5.4 Discussion

In the first project described in this chapter, features derived from data passively collected by a smartphone app were used for predicting decompensation events in a heart failure population. There were three passive data modalities (motion, location, and social contact) and one active (the KCCQ). The hypothesis was that activity, location, and social contact data were all affected by changes in health status for patients with HF, and these data gathered via smartphone could be used to assess patient quality of life. Activity data can be used to infer the circadian rhythm of the patient, activity-rest schedules, and disruptions (such as awakenings during the night). Location data can indicate if the patient is disinclined or unable to leave the house or continue a normal routine. Social contact data provides information about the patient's social interaction, particularly with a given subgroup of contacts. During health changes, we observe a change in social behavior with a user changing the frequency or duration of calls and to whom they are placed. Although none of these changes are particularly specific to changes in HF when taken in isolation, when combined, they provide a more accurate measure of the changes in health. For this dataset, traditional change point detection methods are not feasible for detecting decompensation events since the data is unstructured and the missingness is high.

In the first sub-project, features extracted from the three domains (activity, location, and social contact) were used to build personalized models to estimate the quality of life-related to HF. Personalized models were implemented instead of population models because KCCQ scores are subjective and self-reported. Using only activity data, the MAE in the KCCQ estimate was 5.71%. A K-NN model, which classifies KCCQ scores as $\leq 25$ or $> 25$ was implemented to detect clinically significant changes in the population. This binary classifier exhibited an accuracy of 78% and 88% for the two subjects who had a sufficient number of passive data and KCCQ scores.

Using only passive data sources, as described in this analysis, and eliminating active data uploading can reduce the burden on patients by minimizing their effort required to participate and increase adherence to monitoring. Moreover, the phone app provides a natural communication medium for the caregiver to intervene when significant or sustained deterioration in health is detected. Subjects had the option to determine which types of data to share and the frequency of their uploads. This option was provided to empower users to take ownership of their data, decide when and how they were monitored, and increase compliance. For patients with high mortality risk, these options could be removed to improve data continuity and adherence. It is important to note that patients' adherence to uploading of active data sources can decline rapidly over time [143].

We also observed variation in compliance measured by active data uploading, but this was not a consistently diminishing level as implied in other studies. For example, a subject enrolled for 428 days sent passive activity data for 328 days, although only 46 KCCQ survey reports were completed in the same period. However, we may only need the first few weeks of reports to build individualized models during the higher compliance period. In the cases where we have insufficient data to build a model, or the confidence in the model drops below a given threshold, incentives can be provided to report the KCCQ. In this way, the information is maximized, and the patient burden is minimized.

In the second sub-project, next-day HF decompensation prediction algorithms were built using each modality separately and fusing all data types. From the passive data sources, the motion data-based model achieved the highest AUCPr of 0.60. For a model based only on the responses of the KCCQ, using the summary of all domains and using the most recent score resulted in the best performance with an AUCPr of 0.69 (Table 5.3). Combining both passive and active data modalities achieved a superior performance compared to models based on passive or actively collected data alone (see Table 5.4 and Table 5.5). The highest performing model combined KCCQ, motion, and social contact data. Late fusion approach achieved a 8% higher AUCPr than early fusion when three modalities were used. Late fusion presents a lower-dimensional vector to the final classifier [144]. Therefore, this method could reduce the chances of overfitting and addresses the curse of dimensionality when the sample size is small. The TPR (0.70) and PPV (0.68) of this model could indicate that the approach could potentially add clinical interventions into the framework and result in a low number of false alarms.

Figure 5.11 illustrates the feature importance using the SHAP method. Duration and number of calls were among the most informative features, indicating that the dynamics of social interactions could be affected by the disease status. It can also be seen in the SHAP summary plot that a higher call duration but fewer number of calls result in a higher probability of HF decompensation. Another important feature was the KCCQ summary value, and a lower value of this parameter gave rise to higher SHAP values. The SHAP plot also indicated that higher mean smartphone motion intensity resulted in a higher probability of HF, which was unexpected since HF limits daily physical activity and is often associated with fatigue. In a previous study, Duncan *et al.* have shown that steps measured by a smartphone and a wearable differed a mean bias of 21.5%, and hypothesize that this could result from the behavior of the participants (i.e., not carrying the phone on short walking breaks, carry location for the phone) [145]. Similarly, our results show that the smartphone's motion data does not measure the physical effort but that it reflects patterns of behavior, including phone utilization and body movements.

When different time-to-event horizons were tested, a general trend of lower performance for longer future predictions was observed, as expected, since symptoms are likely to become more pronounced closer to the event. However, predictions two days ahead were better than one day, and performance four days ahead was almost as good as one day before the event. This indicates that one-day, two-day, and four-day models could be run simultaneously to identify short- and medium-term risks and result in different levels of intervention. Changes in performance will be affected by the levels of missingness as the event approached and the intrinsic behaviors, which may explain the performance of the two-day window.

There are two fundamental limitations of the analysis on this project. Firstly, when the data were missing, the app did not indicate whether this resulted from the participant closing the app voluntarily or if it resulted from the smartphone battery running out. These behaviors have different etiologies, which may be related to impending decompensation in different ways. For example, closing the app may indicate being tired, whereas a battery running out of charge may indicate apathy connected with depression. If an additional label is collected for missing sections, it could be used to learn other behavioral patterns. Secondly, even though each participant contributed many days, the study's sample size was relatively small (N=28 participants), and therefore, the methods should be further validated in a larger cohort.

Finally, motion data collected with wearables were investigated for chronic DMD and PTSD cohorts. While cross-validated models could not be built due to small set of data samples, the motion features were compared with clinical outcomes. DMD study indicated that wearable-based approaches could provide an opportunity to monitor DMD patients non-invasively. First, ambulatory participants with more activity rhythm fragmentation had higher ratings of subjective impairment in initiating and maintaining sleep. IV and the DIMS subscale measure overlapping constructs of sleep disruption, and their close association provides evidence that IV is a clinically meaningful indicator of sleep-wake dysfunction in DMD. Second, habitual day-time activity was associated with walk test performance, providing evidence that M10 may be a valid marker of ambulatory status complementary to the 6MWT. Similarly, in the Emory PTSD dataset, $IS$ and $CRS_{\sigma}$ features were statistically significant across healthy and chronic PTSD groups. PTSD group had lower $IS$, which could indicate that this group had lower coupling to environmental zeitgebers that regulate the circadian rhythm.

This group also had a higher and more variant $CRS_\sigma$ feature. Further data collection is required for both of these studies, but the findings in this dissertation could indicate that wearable-based approaches could be useful for passive monitoring at home.

## 5.5 Conclusion

In this chapter, machine learning approaches for monitoring chronic state HF, DMD, and PTSD patients were introduced. Chronic diseases persist over longer periods [129], and could require long-term care to improve the recurring symptoms. Thus, the methods described in this dissertation could be beneficial for patients suffering from chronic conditions and their care providers. Firstly, due to the ubiquity of smartphones or wearables and the ease of scalability of the framework, these methods will facilitate monitoring large populations at a low cost. Secondly, they could be considered minimally invasive and impose less burden on patients already suffering from long-term diseases. In conclusion, the findings from this chapter demonstrate the feasibility of wearable and smartphone-based approaches to monitor chronic diseases and show that these methods could be non-invasive and passive alternatives to existing approaches.

# CHAPTER 6
# CONCLUSION

This dissertation presents wearable-based tools and technologies for improving longitudinal patient monitoring. The overall aim of the dissertation is to present low-cost and non-invasive technology that can be combined with existing tools for more accurate and efficient longitudinal monitoring. All the methods presented in this work are validated using large-scale human-subject studies with data collected in the wild.

The first part of this dissertation focused on building a novel sleep detection model using the physiological and motion signals and discussed how the change points in these signals could be utilized to estimate underlying sleep/wake states. This work demonstrated that only the information in the change points is sufficient to achieve unbiased performance. Storing the signals as change-event time series could reduce the required memory of the device. In this way, this work would enable the memory-constraint wearables to monitor longer durations in-home settings. Furthermore, different smartwatch brands could provide signals with varying amplitudes, but the proposed method was dependent only on the changes in the signals, so it could be well-suited to work on any device.

In the second part of the dissertation, longitudinal wearable-based monitoring techniques are presented for two different applications. The first work showed that these features that quantify the physiological variability in the weeks post-trauma could provide important information regarding the health conditions in the following months. While the wearable-based models were not very useful for estimating PCL-5 survey outcome, the models showed better performance for distinguishing between healthy participants and participants with multiple symptoms. Furthermore, it is shown that including the wearable-based monitoring could improve upon self-report surveys administrated in ED and could be more helpful in identifying changes in the disease status.

The last part of the dissertation discussed using wearables to monitor participants with chronic health conditions such as heart failure. A smartphone-based framework collected passive and active data types from participants. Novel algorithms were developed for predicting decompensation events, and the predictive performance of each data modality was tested. To the author's knowledge, this work represents the first smartphone-based approach for non-invasive longitudinal monitoring for cardiovascular diseases. Fusion approaches for combining all modalities were evaluated and discussed. Most significantly, it is shown that the inclusion of passive metrics improved the performance, and therefore, this approach could improve upon self-report-based methods for monitoring heart failure patients.

Although the passive-monitoring models developed in this dissertation have shown promise in detecting both developing and chronic health conditions, future studies are required for assessing if they could be coupled with clinical interventions and if they could be deployed on wearable devices to be used during free-living conditions. Possible future directions and improvements could include (i) detecting non-wear regions from wearables and handling the missing data, (ii) investigating circadian variability of the heart rate for mental health disorders, (iii) alerting caregivers using disease severity estimates from the models. In the following section, these research questions and possible future directions are discussed.

## 6.1 Future directions

### 6.1.1 Detecting non-wear regions from wearables and handling the missing data

Non-wear periods are recordings of data when the participant does not wear the device, but data is recorded and not marked as missing. These non-wear regions could be confused with sedentary periods and could lead to biases in circadian metrics and inaccuracies for sleep-period detection. A new dataset should be created to validate algorithms for non-wear region detection since none of the datasets used in this dissertation had labels for this task. Existing approaches in the literature focus on the tilt and the activity intensity measured by accelerometer signals [146, 147], the signal quality of the PPG signal or the number of heartbeats detected in the window could also be used as additional tools. After the non-wear periods are detected, these could also be marked as missing data regions. Performance of different missing data imputation techniques could be tested to see the effect on the final model performance.

### 6.1.2 Investigating circadian variability of the heart rate for mental health disorders

HPA axis and the master biological clock SCN regulate the cortisol hormone release to the body through complex biological pathways [148]. Cortisol production also follows a rhythm under the control of these systems, rising before sleep offset and decreasing through the day. Previous studies have shown that patients with mental health could have abnormal cortisol rhythms, with phase advancements [148, 149]. Since cortisol leads to an increase in the heart rate, the metrics derived from the PPG data could track this rhythm. Other HRV metrics such as *LF*, *HF*, or *RMSSD* could also be monitored daily to quantify the stress response through HRV. It was not possible to monitor the circadian variability of the heart rate in this dissertation due to the high missingness of the data during daytime as shown in Figure 4.3. If another wearable is used to collect higher quality PPG data from both daytime and sedentary regions, circadian variability of the heart rate could become another biomarker to monitor the participants.

### 6.1.3 Alerting caregivers using disease severity estimates from the models

Finally, the methods developed in this dissertation could be coupled with clinical interventions. The model's output could be interpreted as a severity indicator instead of class membership and the framework could alert the caregiver if there is a deterioration. For example, for the AMoSS project, the models could be deployed on smartphones, and the app could deliver alerts to the clinical teams. Based on this, the clinicians could check their patients and deliver timely support and corrective therapies. However, whether this approach could reduce hospitalizations and improve patient's quality of life should be tested in real-life scenarios and further validated.

## 6.2 Final remarks

Recent advances in wearable technology provided a new tracking tools for researchers. The methods described in this dissertation could significantly improve passive wearable-based monitoring and provide insights to patient's daily life behaviours outside the clinic. Sleep diaries could be used to find the sleep periods of patients and then CPD method could be used to objectively track sleep disturbances every night. Biological rhythm metrics from wearables or smartphones could track the patient's recovery after hospital discharge. This dissertation provides a basis for a non-invasive health monitoring on multiple scales and these technological advances could become a complementary tool for clinicians.

# REFERENCES

[1] Y.-F. Guo and P. K. Stein, "Circadian rhythm in the cardiovascular system: Chronocardiology," *American Heart Journal*, vol. 145, no. 5, pp. 779–786, 2003.

[2] W. H. Walker, J. C. Walton, A. C. DeVries, and R. J. Nelson, "Circadian rhythm disruption and mental health," *Translational Psychiatry*, vol. 10, no. 1, pp. 1–13, 2020.

[3] A. Germain and D. J. Kupfer, "Circadian rhythm disturbances in depression," *Human Psychopharmacology: Clinical and Experimental*, vol. 23, no. 7, pp. 571–585, 2008.

[4] A. Richards, J. C. Kanady, and T. C. Neylan, "Sleep disturbance in PTSD and other anxiety-related disorders: an updated review of clinical features, physiological characteristics, and psychological and neurobiological mechanisms," *Neuropsychopharmacology*, vol. 45, no. 1, pp. 55–73, 2020.

[5] I. S. Khawaja, J. J. Westermeyer, and T. D. Hurwitz, "Actigraphy and PTSD," in *Sleep and Combat-Related Post Traumatic Stress Disorder*, Springer, 2018, pp. 209–213.

[6] J. Dayan, G. Rauchs, and B. Guillery-Girard, "Rhythms dysregulation: A new perspective for understanding PTSD?" *Journal of Physiology-Paris*, vol. 110, no. 4, pp. 453–460, 2016.

[7] K. S. Gilbert, S. M. Kark, P. Gehrman, and Y. Bogdanova, "Sleep disturbances, TBI and PTSD: implications for treatment and recovery," *Clinical Psychology Review*, vol. 40, pp. 195–212, 2015.

[8] R. Laje, P. V. Agostino, and D. A. Golombek, "The times of our lives: Interaction among different biological periodicities," *Frontiers in Integrative Neuroscience*, vol. 12, p. 10, 2018.

[9] W. A. Hofstra and A. W. de Weerd, "How to assess circadian rhythm in humans: A review of literature," *Epilepsy & Behavior*, vol. 13, no. 3, pp. 438–444, 2008.

[10] M. Gerkema, "Ultradian rhythms," in *Biological Rhythms*, Springer, 2002, pp. 207–215.

[11] P. K. Stein, E. J. Lundequam, L. P. Oliveira, D. J. Clauw, K. E. Freedland, R. M. Carney, and P. P. Domitrovich, "Cardiac autonomic modulation,"

*IEEE Engineering In Medicine And Biology Magazine*, vol. 26, no. 6, p. 14, 2007.

[12] V. Vaccarino, J. Goldberg, C. Rooks, A. J. Shah, E. Veledar, T. L. Faber, J. R. Votaw, C. W. Forsberg, and J. D. Bremner, "Post-traumatic stress disorder and incidence of coronary heart disease: A twin study," *Journal of the American College of Cardiology*, vol. 62, no. 11, pp. 970–978, 2013.

[13] 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.

[14] J. B. Webster, D. F. Kripke, S. Messin, D. J. Mullaney, and G. Wyborney, "An activity-based sleep monitor system for ambulatory use," *Sleep*, vol. 5, no. 4, pp. 389–399, 1982.

[15] J. Lötjönen, I. Korhonen, K. Hirvonen, S. Eskelinen, M. Myllymäki, and M. Partinen, "Automatic sleep-wake and nap analysis with a new wrist worn online activity monitoring device vivago WristCare®," *Sleep*, vol. 26, no. 1, pp. 86–90, 2003.

[16] J. Hedner, G. Pillar, S. D. Pittman, D. Zou, L. Grote, and D. P. White, "A novel adaptive wrist actigraphy algorithm for sleep-wake assessment in sleep apnea patients," *Sleep*, vol. 27, no. 8, pp. 1560–1566, 2004.

[17] A. Goldstone, F. C. Baker, and M. de Zambotti, "Actigraphy in the digital health revolution: Still asleep?" *Sleep*, vol. 41, no. 9, zsy120, 2018.

[18] M. Marino, Y. Li, M. N. Rueschman, J. W. Winkelman, J. Ellenbogen, J. M. Solet, H. Dulin, L. F. Berkman, and O. M. Buxton, "Measuring sleep: Accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography," *Sleep*, vol. 36, no. 11, pp. 1747–1755, 2013.

[19] J. Paquet, A. Kawinska, and J. Carrier, "Wake detection capacity of actigraphy during sleep," *Sleep*, vol. 30, no. 10, pp. 1362–1369, 2007.

[20] N. Oakley, "Validation with polysomnography of the Sleepwatch sleep/wake scoring algorithm used by the actiwatch activity monitoring system," *Bend: Mini Mitter, Cambridge Neurotechnology*, 1997.

[21] B. Sivertsen, S. Omvik, O. E. Havik, S. Pallesen, B. Bjorvatn, G. H. Nielsen, S. Straume, and I. H. Nordhus, "A comparison of actigraphy and polysomnography in older adults treated for chronic primary insomnia," *Sleep*, vol. 29, no. 10, pp. 1353–1358, 2006.

[22] S.-G. Kang, J. M. Kang, K.-P. Ko, S.-C. Park, S. Mariani, and J. Weng, "Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers," *Journal of Psychosomatic Research*, vol. 97, pp. 38–44, 2017.

[23] F. Chouchou and M. Desseilles, "Heart rate variability: A tool to explore the sleeping brain?" *Frontiers in Neuroscience*, vol. 8, p. 402, 2014.

[24] G. Varoneckas, K. Plauška, J. Kauk, *et al.*, "Components of the heart rhythm power spectrum in wakefulness and individual sleep stages," *International Journal of Psychophysiology*, vol. 4, no. 2, pp. 129–141, 1986.

[25] M. Bonnet and D. Arand, "Heart rate variability: Sleep stage, time of night, and arousal influences," *Electroencephalography and Clinical Neurophysiology*, vol. 102, no. 5, pp. 390–396, 1997.

[26] P. Fonseca, T. Weysen, M. S. Goelema, E. I. Møst, M. Radha, C. Lunsingh Scheurleer, L. van den Heuvel, and R. M. Aarts, "Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults," *Sleep*, vol. 40, no. 7, zsx097, 2017.

[27] Z. Beattie, Y. Oyang, A. Statan, A. Ghoreyshi, A. Pantelopoulos, A. Russell, and C. Heneghan, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiological Measurement*, vol. 38, no. 11, p. 1968, 2017.

[28] S. Eyal and A. Baharav, "Sleep insights from the finger tip: How photoplethysmography can help quantify sleep," in *2017 Computing in Cardiology (CinC)*, IEEE, 2017, pp. 1–4.

[29] J. I. Bisson, "Post-traumatic stress disorder," *BMJ*, vol. 334, no. 7597, pp. 789–793, 2007.

[30] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[31] T. J. Sharp and A. G. Harvey, "Chronic pain and posttraumatic stress disorder: Mutual maintenance?" *Clinical Psychology Review*, vol. 21, no. 6, pp. 857–877, 2001.

[32] D. A. Fishbain, A. Pulikal, J. E. Lewis, and J. Gao, "Chronic pain types differ in their reported prevalence of post-traumatic stress disorder (PTSD) and there is consistent evidence that chronic pain is associated with PTSD: an evidence-based structured systematic review," *Pain Medicine*, vol. 18, no. 4, pp. 711–735, 2017.

[33] D. G. Kilpatrick, H. S. Resnick, M. E. Milanak, M. W. Miller, K. M. Keyes, and M. J. Friedman, "National estimates of exposure to traumatic events and PTSD prevalence using DSM-IV and DSM-5 criteria," *Journal of Traumatic Stress*, vol. 26, no. 5, pp. 537–547, 2013.

[34] B. Andrews, C. R. Brewin, R. Philpott, and L. Stewart, "Delayed-onset posttraumatic stress disorder: A systematic review of the evidence," *American Journal of Psychiatry*, vol. 164, no. 9, pp. 1319–1326, 2007.

[35] E. J. Ozer, S. R. Best, T. L. Lipsey, and D. S. Weiss, "Predictors of posttraumatic stress disorder and symptoms in adults: A meta-analysis.," *Psychological Bulletin*, vol. 129, no. 1, p. 52, 2003.

[36] National Collaborating Centre for Mental Health (UK and others), "Predictors of PTSD and screening for the disorder," in *Post-Traumatic Stress Disorder: The Management of PTSD in Adults and Children in Primary and Secondary Care*, Gaskell, 2005.

[37] C. R. Brewin, B. Andrews, and J. D. Valentine, "Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults.," *Journal of Consulting and Clinical Psychology*, vol. 68, no. 5, p. 748, 2000.

[38] K. Schultebraucks, A. Y. Shalev, V. Michopoulos, C. R. Grudzen, S.-M. Shin, J. S. Stevens, J. L. Maples-Keller, T. Jovanovic, G. A. Bonanno, B. O. Rothbaum, *et al.*, "A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor," *Nature Medicine*, vol. 26, no. 7, pp. 1084–1088, 2020.

[39] K. Schultebraucks, V. Yadav, A. Y. Shalev, G. A. Bonanno, and I. R. Galatzer-Levy, "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," *Psychological Medicine*, pp. 1–11, 2020.

[40] E. Reinertsen and G. D. Clifford, "A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses," *Physiological Measurement*, vol. 39, no. 5, 05TR01, 2018.

[41] G. Valenza, M. Nardelli, A. Lanata, C. Gentili, G. Bertschy, R. Paradiso, and E. P. Scilingo, "Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1625–1635, 2013.

[42] R. M. Carney, K. E. Freedland, P. K. Stein, G. E. Miller, B. Steinmeyer, M. W. Rich, and S. P. Duntley, "Heart rate variability and markers of inflamma-

tion and coagulation in depressed patients with coronary heart disease," *Journal of Psychosomatic Research*, vol. 62, no. 4, pp. 463–467, 2007.

[43] J. M. Dekker, R. S. Crow, A. R. Folsom, P. J. Hannan, D. Liao, C. A. Swenne, and E. G. Schouten, "Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: the ARIC Study," *Circulation*, vol. 102, no. 11, pp. 1239–1244, 2000.

[44] T. A. Mellman, B. R. Knorr, W. R. Pigeon, J. Leiter, and M. Akay, "Heart rate variability during sleep and the early development of posttraumatic stress disorder," *Biological Psychiatry*, vol. 55, no. 9, pp. 953–956, 2004.

[45] G. J. van Boxtel, P. J. Cluitmans, R. J. Raymann, M. Ouwerkerk, A. J. Denissen, M. K. Dekker, and M. M. Sitskoorn, "Heart rate variability, sleep, and the early detection of post-traumatic stress disorder," in *Sleep and Combat-Related Post-traumatic Stress Disorder*, Springer, 2018, pp. 253–263.

[46] E. Reinertsen, S. Nemati, A. N. Vest, V. Vaccarino, R. Lampert, A. J. Shah, and G. D. Clifford, "Heart rate-based window segmentation improves accuracy of classifying posttraumatic stress disorder using heart rate variability measures," *Physiological Measurement*, vol. 38, no. 6, p. 1061, 2017.

[47] A. D. McDonald, F. Sasangohar, A. Jatav, and A. H. Rao, "Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: a supervised machine learning approach," *IISE Transactions on Healthcare Systems Engineering*, vol. 9, no. 3, pp. 201–211, 2019.

[48] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," *Sleep*, vol. 26, no. 3, pp. 342–392, 2003.

[49] D. J. Inman, S. M. Silver, and K. Doghramji, "Sleep disturbance in post-traumatic stress disorder: a comparison with non-PTSD insomnia," *Journal of Traumatic Stress*, vol. 3, no. 3, pp. 429–437, 1990.

[50] A. Tsanas, E. Woodward, and A. Ehlers, "Objective characterization of activity, sleep, and circadian rhythm patterns using a wrist-worn actigraphy sensor: Insights into posttraumatic stress disorder," *JMIR mHealth and uHealth*, vol. 8, no. 4, e14306, 2020.

[51] S. S. Virani, A. Alonso, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, F. N. Delling, *et al.*, "Heart disease and stroke statistics—2020 update: a report from the American Heart Association," *Circulation*, E139–E596, 2020.

[52] S. Chen, M. Kuhn, K. Prettner, and D. E. Bloom, "The macroeconomic burden of noncommunicable diseases in the United States: Estimates and projections," *PloS One*, vol. 13, no. 11, e0206702, 2018.

[53] G. M. Felker, K. F. Adams Jr, M. A. Konstam, C. M. O'Connor, and M. Gheorghiade, "The problem of decompensated heart failure: Nomenclature, classification, and risk stratification," *American Heart Journal*, vol. 145, no. 2, S18–S25, 2003.

[54] S. M. Joseph, A. M. Cedars, G. A. Ewald, E. M. Geltman, and D. L. Mann, "Acute decompensated heart failure: Contemporary medical management," *Texas Heart Institute Journal*, vol. 36, no. 6, p. 510, 2009.

[55] S. Patil, M. Shah, B. Patel, M. Agarwal, P. Ram, and V. M. Alla, "Readmissions among patients admitted with acute decompensated heart failure based on income quartiles," in *Mayo Clinic Proceedings*, Elsevier, vol. 94, 2019, pp. 1939–1950.

[56] M. Packer, W. T. Abraham, M. R. Mehra, C. W. Yancy, C. E. Lawless, J. E. Mitchell, F. W. Smart, R. Bijou, C. M. O'Connor, B. M. Massie, *et al.*, "Utility of impedance cardiography for the identification of short-term risk of clinical decompensation in stable patients with chronic heart failure," *Journal of the American College of Cardiology*, vol. 47, no. 11, pp. 2245–2252, 2006.

[57] I. S. Anand, W. W. Tang, B. H. Greenberg, N. Chakravarthy, I. Libbus, R. P. Katra, M. Investigators, *et al.*, "Design and performance of a multisensor heart failure monitoring algorithm: results from the multisensor monitoring in congestive heart failure (MUSIC) study," *Journal of Cardiac Failure*, vol. 18, no. 4, pp. 289–295, 2012.

[58] O. T. Inan, M. Baran Pouyan, A. Q. Javaid, S. Dowling, M. Etemadi, A. Dorier, J. A. Heller, A. O. Bicen, S. Roy, T. De Marco, *et al.*, "Novel wearable seismocardiography and machine learning algorithms can assess clinical status of heart failure patients," *Circulation: Heart Failure*, vol. 11, no. 1, e004313, 2018.

[59] J. Stehlik, C. Schmalfuss, B. Bozkurt, J. Nativi-Nicolau, P. Wohlfahrt, S. Wegerich, K. Rose, R. Ray, R. Schofield, A. Deswal, *et al.*, "Continuous wearable monitoring analytics predict heart failure hospitalization: The LINK-HF multicenter study," *Circulation: Heart Failure*, vol. 13, no. 3, e006513, 2020.

[60] V. B. Aydemir, S. Nagesh, M. M. H. Shandhi, J. Fan, L. Klein, M. Etemadi, J. A. Heller, O. T. Inan, and J. M. Rehg, "Classification of decompensated

heart failure from clinical and home ballistocardiography," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 5, pp. 1303–1313, 2019.

[61] C. P. Green, C. B. Porter, D. R. Bresnahan, and J. A. Spertus, "Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure," *Journal of the American College of Cardiology*, vol. 35, no. 5, pp. 1245–1255, 2000.

[62] K. E. Flynn, L. Lin, S. J. Ellis, S. D. Russell, J. A. Spertus, D. J. Whellan, I. L. Piña, L. J. Fine, K. A. Schulman, and K. P. Weinfurt, "Relationships between patient-reported outcome measures and clinical measures in outpatients with heart failure," *American Heart Journal*, vol. 158, no. 4 Suppl, S64, 2009.

[63] N. Palmius, M. Osipov, A. Bilderbeck, G. Goodwin, K. Saunders, A. Tsanas, and G. Clifford, "A multi-sensor monitoring system for objective mental health management in resource constrained environments," in *Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*.

[64] A. S. Cakmak, E. Reinertsen, H. A. Taylor, A. J. Shah, and G. D. Clifford, "Personalized heart failure severity estimates using passive smartphone data," in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 1569–1574.

[65] A. S. Cakmak, H. J. Lanier, E. Reinertsen, A. Harzand, A. M. Zafari, M. A. Hammoud, A. Alrohaibani, C. Wakwe, M. Appeadu, G. D. Clifford, *et al.*, "Passive smartphone actigraphy data predicts heart failure decompensation," *Circulation*, vol. 140, no. Suppl_1, A15444–A15444, 2019.

[66] A. S. Cakmak, S. Densen, G. Najarro, P. Rout, C. J. Rozell, O. T. Inan, A. J. Shah, and G. D. Clifford, "Late fusion of machine learning models using passively captured interpersonal social interactions and motion from smartphones predicts decompensation in heart failure," *arXiv preprint arXiv:2104.01511*, 2021.

[67] S. S. Coughlin, "Post-traumatic stress disorder and cardiovascular disease," *The Open Cardiovascular Medicine Journal*, vol. 5, p. 164, 2011.

[68] B. A. Wentworth, M. B. Stein, L. S. Redwine, Y. Xue, P. R. Taub, P. Clopton, K. R. Nayak, and A. S. Maisel, "Post-traumatic stress disorder: A fast track to premature cardiovascular disease?" *Cardiology in Review*, vol. 21, no. 1, pp. 16–22, 2013.

[69] H. K. Kang, T. A. Bullman, and J. W. Taylor, "Risk of selected cardiovascular diseases and posttraumatic stress disorder among former World War II prisoners of war," *Annals of Epidemiology*, vol. 16, no. 5, pp. 381–386, 2006.

[70] E. M. Yiu, A. J. Kornberg, *et al.*, “Duchenne muscular dystrophy,” *Neurology India*, vol. 56, no. 3, p. 236, 2008.

[71] C. Bloetzer, P.-Y. Jeannet, B. Lynch, and C. J. Newman, “Sleep disorders in boys with Duchenne muscular dystrophy,” *Acta Paediatrica*, vol. 101, no. 12, pp. 1265–1269, 2012.

[72] E. M. Yiu and A. J. Kornberg, “Duchenne muscular dystrophy,” *Journal of Paediatrics and Child Health*, vol. 51, no. 8, pp. 759–764, 2015.

[73] Z. E. Davidson, M. M. Ryan, A. J. Kornberg, K. Z. Walker, and H. Truby, “Strong correlation between the 6-minute walk test and accelerometry functional outcomes in boys with Duchenne muscular dystrophy,” *Journal of Child Neurology*, vol. 30, no. 3, pp. 357–363, 2015.

[74] S. Kimura, S. Ozasa, K. Nomura, K. Yoshioka, and F. Endo, “Estimation of muscle strength from actigraph data in Duchenne muscular dystrophy,” *Pediatrics International*, vol. 56, no. 5, pp. 748–752, 2014.

[75] M. Tsai, A. M. Mori, C. W. Forsberg, N. Waiss, J. L. Sporleder, N. L. Smith, and J. Goldberg, “The Vietnam era twin registry: A quarter century of progress,” *Twin Research and Human Genetics*, vol. 16, no. 1, pp. 429–436, 2013.

[76] S. A. McLean, K. Ressler, K. C. Koenen, T. Neylan, L. Germine, T. Jovanovic, G. D. Clifford, D. Zeng, X. An, S. Linnstaedt, *et al.*, “The AURORA Study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure,” *Molecular Psychiatry*, vol. 25, no. 2, pp. 283–296, 2020.

[77] N. Palmius, M. Osipov, A. C. Bilderbeck, G. M. Goodwin, K. Saunders, A. Tsanas, and G. D. Clifford, “A multi-sensor monitoring system for objective mental health management in resource constrained environments,” in *Appropriate Healthcare Technologies for Low Resource Settings*, 2014.

[78] B. L. Lan, E. V. Yeoh, and J. A. Ng, “Distribution of detrended stock market data,” *Fluctuation and Noise Letters*, vol. 9, no. 03, pp. 245–257, 2010.

[79] Z. Wu, N. E. Huang, S. R. Long, and C.-K. Peng, “On the trend, detrending, and variability of nonlinear and nonstationary time series.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 38, pp. 14 889–94, 2007.

[80] L. E. Raffalovich, “Detrending time series: A cautionary note,” *Sociological Methods & Research*, vol. 22, no. 4, pp. 492–519, May 1994.

[81] C. R. Nelson and H. Kang, "Spurious periodicity in inappropriately detrended time series," *Econometrica*, vol. 49, no. 3, pp. 741–51, 1981.

[82] P. Bernaola-Galván, P. Ivanov, L. Nunes Amaral, and H. Stanley, "Scale invariance in the nonstationarity of human heart rate," *Physical Review Letters*, vol. 87, no. 16, p. 168 105, Oct. 2001.

[83] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, "Method for generating an artificial RR tachogram of a typical healthy human over 24-hours," in *Computers in Cardiology*, IEEE, 2002, pp. 225–228.

[84] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, Physiotoolkit, and Physionet," *Circulation*, vol. 101, no. 23, e215–e220, 2000.

[85] P. Bernaola-Galván, J. L. Oliver, M. Hackenberg, A. V. Coronado, P. C. Ivanov, and P. Carpena, "Segmentation of time series with long-range fractal correlations.," *The European Physical Journal. B*, vol. 85, no. 6, Jun. 2012.

[86] K. Fukuda, H. E. Stanley, and L. A. N. Amaral, "Heuristic segmentation of a nonstationary time series," *Physical Review E*, vol. 69, no. 2, p. 021 108, 2004.

[87] J. D. Scargle, J. P. Norris, B. Jackson, and J. Chiang, "Studies in astronomical time series analysis. VI. Bayesian block representations," *The Astrophysical Journal*, vol. 764, no. 2, p. 167, Feb. 2013.

[88] J. Chen and A. K. Gupta, *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media, 2011.

[89] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[90] R. P. Adams and D. J. C. MacKay, "Bayesian Online Changepoint Detection," University of Cambridge, Cambridge, UK, Tech. Rep., 2007.

[91] R. Turner, Y. Saatci, and C. E. Rasmussen, *Adaptive Sequential Bayesian Change Point Detection*, Dec. 2009.

[92] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," *DEF*, vol. 1, no. $2\sigma2$, p. 16, 2007.

[93] A. S. Cakmak, G. Da Poian, A. Willats, A. Haffar, R. Abdulbaki, Y.-A. Ko, A. J. Shah, V. Vaccarino, D. L. Bliwise, C. Rozell, *et al.*, "An unbiased, efficient sleep–wake detection algorithm for a population with sleep disorders: Change point decoder," *Sleep*, 2020.

[94] J. W. Pillow, Y. Ahmadian, and L. Paninski, "Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains," *Neural Computation*, vol. 23, no. 1, pp. 1–45, 2011.

[95] A. N. Vest, G. Da Poian, Q. Li, C. Liu, S. Nemati, A. J. Shah, and G. D. Clifford, "An open source benchmarked toolbox for cardiovascular waveform and interval analysis," *Physiological Measurement*, vol. 39, no. 10, p. 105 004, 2018.

[96] G. Welch, G. Bishop, *et al.*, "An introduction to the Kalman filter," 1995.

[97] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiological Measurement*, vol. 29, no. 1, p. 15, 2007.

[98] M. Borazio, E. Berlin, N. Kucukyildiz, P. Scholl, and K. V. Laerhoven, "Towards benchmarked sleep detection with inertial wrist-worn sensing units," *IEEE International Conference on Healthcare Informatics*, 2014.

[99] M. Yoneyama, Y. Okuma, H. Utsumi, H. Terashi, and H. Mitoma, "Human turnover dynamics during sleep: Statistical behavior and its modeling," *Physical Review E*, vol. 89, no. 3, p. 032 721, 2014.

[100] A. Kosmadopoulos, C. Sargent, D. Darwent, X. Zhou, and G. D. Roach, "Alternatives to polysomnography (PSG): a validation of wrist actigraphy and a partial-PSG system," *Behavior Research Methods*, vol. 46, no. 4, pp. 1032–1041, 2014.

[101] K. Lichstein, H. Durrence, D. Taylor, A. Bush, and B. Riedel, "Quantitative criteria for insomnia," *Behaviour Research and Therapy*, vol. 41, no. 4, pp. 427–445, 2003.

[102] G. Cornelissen, "Cosinor-based rhythmometry," *Theoretical Biology and Medical Modelling*, vol. 11, no. 1, pp. 1–24, 2014.

[103] A. Michaels, C. Michaels, C. Moon, M. A. Zimmerman, C. Peterson, and J. L. Rodriguez, "Psychosocial factors limit outcomes after trauma," *Journal of Trauma and Acute Care Surgery*, vol. 44, no. 4, pp. 644–648, 1998.

[104] A. Brunet, D. S. Weiss, T. J. Metzler, S. R. Best, T. C. Neylan, C. Rogers, J. Fagan, and C. R. Marmar, "The Peritraumatic Distress Inventory: a proposed measure of PTSD criterion A2," *American Journal of Psychiatry*, vol. 158, no. 9, pp. 1480–1485, 2001.

[105] M. J. Bovin, B. P. Marx, F. W. Weathers, M. W. Gallagher, P. Rodriguez, P. P. Schnurr, and T. M. Keane, "Psychometric properties of the PTSD checklist for diagnostic and statistical manual of mental disorders–fifth edition (PCL-5) in veterans.," *Psychological assessment*, vol. 28, no. 11, p. 1379, 2016.

[106] F. W. Weathers, B. T. Litz, T. M. Keane, P. A. Palmieri, B. P. Marx, and P. P. Schnurr, "The PTSD checklist for DSM-5 (PCL-5)," *Scale available from the National Center for PTSD at www. ptsd. va. gov*, vol. 10, 2013.

[107] A. Germain, M. Hall, B. Krakow, M. K. Shear, and D. J. Buysse, "A brief sleep scale for posttraumatic stress disorder: Pittsburgh Sleep Quality Index Addendum for PTSD," *Journal of Anxiety Disorders*, vol. 19, no. 2, pp. 233–244, 2005.

[108] S. P. Insana, M. Hall, D. J. Buysse, and A. Germain, "Validation of the Pittsburgh Sleep quality index addendum for posttraumatic stress disorder (PSQI-A) in US Male military veterans," *Journal of Traumatic Stress*, vol. 26, no. 2, pp. 192–200, 2013.

[109] J. A. Teresi, K. Ocepek-Welikson, K. F. Cook, M. Kleinman, M. Ramirez, M. C. Reid, and A. Siu, "Measurement equivalence of the patient reported outcomes measurement information system®(PROMIS®) pain interference short form items: Application to ethnically diverse cancer and palliative care populations," *Psychological Test and Assessment Modeling*, vol. 58, no. 2, p. 309, 2016.

[110] E. J. van Someren, E. E. Hagebeuk, C. Lijzenga, P. Scheltens, S. E. de Rooij, C. Jonker, A.-M. Pot, M. Mirmiran, and D. F. Swaab, "Circadian rest—activity rhythm disturbances in Alzheimer's disease," *Biological Psychiatry*, vol. 40, no. 4, pp. 259–270, 1996.

[111] E. J. Van Someren, D. F. Swaab, C. C. Colenda, W. Cohen, W. V. McCall, and P. B. Rosenquist, "Bright light therapy: Improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods," *Chronobiology International*, vol. 16, no. 4, pp. 505–518, 1999.

[112] Y.-L. Huang, R.-Y. Liu, Q.-S. Wang, E. J. Van Someren, H. Xu, and J.-N. Zhou, "Age-associated difference in circadian sleep–wake and rest–activity rhythms," *Physiology & Behavior*, vol. 76, no. 4-5, pp. 597–603, 2002.

[113] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: Experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[114] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[115] T. H. Monk, C. F. REYNOLDS III, D. J. Kupfer, D. J. Buysse, P. A. Coble, A. J. Hayes, M. A. Machen, S. R. Petrie, and A. M. Ritenour, "The Pittsburgh sleep diary," *Journal of Sleep Research*, vol. 3, no. 2, pp. 111–120, 1994.

[116] C. M. Morin, G. Belleville, L. Bélanger, and H. Ivers, "The insomnia severity index: Psychometric indicators to detect insomnia cases and evaluate treatment response," *Sleep*, vol. 34, no. 5, pp. 601–608, 2011.

[117] C. A. Blevins, F. W. Weathers, M. T. Davis, T. K. Witte, and J. L. Domino, "The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation," *Journal of Traumatic Stress*, vol. 28, no. 6, pp. 489–498, 2015.

[118] D. Amtmann, J. Kim, H. Chung, A. M. Bamer, R. L. Askew, S. Wu, K. F. Cook, and K. L. Johnson, "Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis.," *Rehabilitation Psychology*, vol. 59, no. 2, p. 220, 2014.

[119] J. E. Ware Jr, M. Kosinski, and S. D. Keller, "A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity," *Medical Care*, pp. 220–233, 1996.

[120] S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. Berger, and R. J. Cohen, "Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control," *Science*, vol. 213, no. 4504, pp. 220–222, 1981.

[121] P. A. Dennis, L. Watkins, P. S. Calhoun, A. Oddone, A. Sherwood, M. F. Dennis, M. B. Rissling, and J. C. Beckham, "Posttraumatic stress, heart-rate variability, and the mediating role of behavioral health risks," *Psychosomatic Medicine*, vol. 76, no. 8, p. 629, 2014.

[122] D. McDuff, S. Gontarek, and R. Picard, "Remote measurement of cognitive stress via heart rate variability," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2014, pp. 2957–2960.

[123] A. Hernando, J. Lazaro, E. Gil, A. Arza, J. M. Garzón, R. Lopez-Anton, C. de la Camara, P. Laguna, J. Aguiló, and R. Bailón, "Inclusion of respiratory frequency information in heart rate variability analysis for stress assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1016–1025, 2016.

[124] E. J. W. V. Someren, D. F. Swaab, C. C. Colenda, W. Cohen, W. V. Mccall, P. B. Rosenquist, W. V. Someren, E. J. W. Van Someren, : W. V. Mccall, and P. B. Rosenquist3, "Bright light therapy: Improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods," *Chronobiology International*, vol. 16, no. 164, pp. 505–518, 1999.

[125] Y. Dagan, Y. Zinger, and P. Lavie, "Actigraphic sleep monitoring in posttraumatic stress disorder (PTSD) patients," *Journal of Psychosomatic Research*, vol. 42, no. 6, pp. 577–581, 1997.

[126] M. C. Kearns, K. J. Ressler, D. Zatzick, and B. O. Rothbaum, "Early interventions for PTSD: a review," *Depression and Anxiety*, vol. 29, no. 10, pp. 833–842, 2012.

[127] B. O. Rothbaum, M. C. Kearns, M. Price, E. Malcoun, M. Davis, K. J. Ressler, D. Lang, and D. Houry, "Early intervention may prevent the development of posttraumatic stress disorder: A randomized pilot civilian study with modified prolonged exposure," *Biological Psychiatry*, vol. 72, no. 11, pp. 957–963, 2012.

[128] J. C. Shipherd, M. Keyes, T. Jovanovic, D. J. Ready, D. Baltzell, V. Worley, V. Gordon-Brown, C. Hayslett, and E. Duncan, "Veterans seeking treatment for posttraumatic stress disorder: What about comorbid chronic pain?" *Journal of Rehabilitation Research & Development*, vol. 44, no. 2, 2007.

[129] S. Bernell and S. W. Howard, "Use your words carefully: What is a chronic disease?" *Frontiers in Public Health*, vol. 4, p. 159, 2016.

[130] A. S. Cakmak, P. B. Suresha, and G. D. Clifford, *Open source actigraphy toolbox*, https://doi.org/10.5281/zenodo.4287769,2020.

[131] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," *Sleep*, vol. 26, no. 3, pp. 342–392, 2003.

[132] P. Jones, K. Gosch, Y. Li, K. Reid, F. Tang, P. Chan, and J. Spertus, "The KCCQ-12: A short version of the Kansas City Cardiomyopathy Questionnaire," *Circulation: Cardiovascular Quality and Outcomes*, vol. 6, no. Suppl 1, 2013.

[133] M. C. Mormont, J. Waterhouse, P. Bleuzen, S. Giacchetti, A. Jami, A. Bogdan, J. Lellouch, J. L. Misset, Y. Touitou, and F. Levi, "Marked 24-h rest/activity rhythms are associated with better quality of life, better response, and longer survival in patients with metastatic colorectal cancer and good performance status," *Clinical Cancer Research*, vol. 6, no. 8, pp. 3038–45, 2000.

[134] M. Costa, A. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, 2002.

[135] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[136] C. P. Green, C. B. Porter, D. R. Bresnahan, and J. A. Spertus, "Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure," *Journal of the American College of Cardiology*, vol. 35, no. 5, pp. 1245–1255, 2000.

[137] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[138] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[139] B. I. Siegel, A. Cakmak, E. Reinertsen, M. Benoit, J. Figueroa, G. D. Clifford, and H. C. Phan, "Use of a wearable device to assess sleep and motor function in Duchenne muscular dystrophy," *Muscle & Nerve*, vol. 61, no. 2, pp. 198–204, 2020.

[140] C. M. McDonald, E. K. Henricson, J. J. Han, R. T. Abresch, A. Nicorici, G. L. Elfring, L. Atkinson, A. Reha, S. Hirawat, and L. L. Miller, "The 6-minute walk test as a new outcome measure in Duchenne muscular dystrophy," *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, vol. 41, no. 4, pp. 500–510, 2010.

[141] R. Geiger, A. Strasak, B. Treml, K. Gasser, A. Kleinsasser, V. Fischer, H. Geiger, A. Loeckinger, and J. I. Stein, "Six-minute walk test in children and adolescents," *The Journal of Pediatrics*, vol. 150, no. 4, pp. 395–399, 2007.

[142] O. Bruni, S. Ottaviano, V. Guidetti, M. Romoli, M. Innocenzi, F. Cortesi, and F. Giannotti, "The Sleep Disturbance Scale for Children (SDSC) Construct ion and validation of an instrument to evaluate sleep disturbances in childhood and adolescence," *Journal of Sleep Research*, vol. 5, no. 4, pp. 251–261, 1996.

[143] W. Speier, E. Dzubur, M. Zide, C. Shufelt, S. Joung, J. E. Van Eyk, C. N. Bairey Merz, M. Lopez, B. Spiegel, and C. Arnold, "Evaluating utility and compliance in a patient-based eHealth study using continuous-time heart rate and activity trackers," *Journal of the American Medical Informatics Association*, 2018.

[144] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[145] M. J. Duncan, K. Wunderlich, Y. Zhao, and G. Faulkner, "Walk this way: validity evidence of iPhone health application step count in laboratory and free-living conditions," *Journal of Sports Sciences*, vol. 36, no. 15, pp. 1695–1704, 2018.

[146] M. N. Ahmadi, N. Nathan, R. Sutherland, L. Wolfenden, and S. G. Trost, "Non-wear or sleep? evaluation of five non-wear detection algorithms for raw accelerometer data," *Journal of sports sciences*, vol. 38, no. 4, pp. 399–404, 2020.

[147] L. Jaeschke, A. Luzak, A. Steinbrecher, S. Jeran, M. Ferland, B. Linkohr, H. Schulz, and T. Pischon, "24 h-accelerometry in epidemiological studies: Automated detection of non-wear time in comparison to diary information," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[148] N. Gotlieb, J. Moeller, and L. J. Kriegsfeld, "Circadian control of neuroendocrine function: Implications for health and disease," *Current Opinion in Physiology*, vol. 5, pp. 133–140, 2018.

[149] P. Linkowski, J. Mendlewicz, M. Kerkhofs, R. Leclercq, J. GOLSTEIN, M. BRASSEUR, G. Copinschi, and E. V. CAUTER, "24-hour profiles of adrenocorticotropin, cortisol, and growth hormone in major depressive illness: Effect of antidepressant treatment," *The Journal of Clinical Endocrinology & Metabolism*, vol. 65, no. 1, pp. 141–152, 1987.