Summer 8-15-2022

Smart Sensing and Clinical Predictions with Wearables: From Physiological Signals to Mental Health

Ruixuan Dai
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Computer Science and Engineering

Dissertation Examination Committee:
Chenyang Lu, Chair
Tao Ju
Thomas Kannampallil
Neal Patwari
Ning Zhang

Smart Sensing and Clinical Predictions with Wearables: From Physiological Signals to Mental Health

by
Ruixuan Dai

A dissertation presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2022
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to sincerely thank my advisor, Prof. Chenyang Lu, for his support and guidance throughout my Ph.D. life. His professional expertise and insightful advice were invaluable in formulating my research, shaping my critical thinking, and developing my soft skills. Those intangible properties will be my life-long treasures to guide my future success. Meanwhile, allow me to express my great gratitude to his family including Prof. Lan Yang and all family members for their caring and help.

Also, I would like to greatly thank Prof. Thomas Kannampallil, who provides me with enormous assistance in my research. His enthusiasm for exploring new research questions facilitates me to deliver the insights presented in this dissertation. I will miss every light-hearted meeting with him.

I felt grateful to conduct my Ph.D. research. I would like to thank Prof. Jun Ma for the joyful and fruitful collaboration on the mental health study, and my other thesis committee members, Prof. Ning Zhang, Prof. Tao Ju, and Prof. Neal Patwari for their constructive comments on my dissertation. My thanks must also go to my friends and CPSL members: Dr. Jing Li, Dr. Chong Li, Dr. Yehan Ma, Dr. Haoran Li, Dr. Dingwen Li, Bing Xue, Jingwen Zhang, Hanyang Liu, Ruiqi Wang, and Moran Xu, for their company during my Ph.D. journey.

Last but not least, I would like to thank my wife, who is the best partner and brought me my loveliest daughter during my Ph.D. journey. I sincerely thank my parents and parents-in-law, who always stand by me and support me materially and spiritually.

Ruixuan Dai

Washington University in St. Louis

August 2022

Dedicated to my family.

ABSTRACT OF THE DISSERTATION

Smart Sensing and Clinical Predictions with Wearables:

From Physiological Signals to Mental Health

by

Ruixuan Dai

Doctor of Philosophy in Computer Science

Washington University in St. Louis, 2022

Professor Chenyang Lu, Chair

Wearable devices such as smartwatches and wristbands are gaining adoption. Recent advances in technology in wearables enable remote health monitoring. However, there are challenges in exploiting wearables in healthcare applications. First, sensor readings from wearables are vulnerable to motion and noise artifacts. A robust pipeline is needed to extract reliable measurements from noisy signals. Second, while wearables support an increasing number of sensing modalities, there is a significant need to generate more clinically meaningful measurements with wearables. Finally, to incorporate wearables into clinical practice, we need to establish the link between wearable measurements and clinical outcomes, thus supporting clinical decisions. To facilitate applications of wearables in healthcare, this dissertation research exploits wearables to predict a wide range of clinically relevant outcomes from physiological measurements to mental health disorders:

Measuring respiratory rate on a smartwatch with photoplethysmography: Modern smartwatches usually lack the ability to accurately measure the respiratory rate (RR) in ambulatory settings. We presented the *RespWatch*, an application that can robustly measure RR and run completely on the smartwatch hardware. RespWatch directly reads the PPG waveform from the smartwatch, and utilizes a hybrid approach with both signal processing and deep

learning techniques to handle the noisy sensor signals and generate robust RR measurements. A user study involving various activities showed that our hybrid method has advantages of both accuracy and efficiency over the previous approaches.

Detecting objective and subjective stress with a commercial smartwatch: In this work, we built stress detection models with commercial smartwatches, and we compared the objective stress detection (based on the objective marker of stressor tasks) with the subjective stress detection (based on the user's subjective responses). Results showed that the generic subjective stress models have worse performance than the objective stress models. To enhance the subjective stress detection, we proposed a personalized subjective model accounting for inter-individual differences via adaptive thresholds. Our personalized approach demonstrated better performance.

Multi-task learning for randomized controlled trials (RCTs) with wearables: In this work, we exploited machine learning models in conjunction with RCTs for personalized predictions of a depression treatment outcome in which patients were monitored with wearables. We formulated the predictions in different groups from an RCT as a multi-task learning (MTL) problem, and proposed a novel MTL model specifically designed for the RCT. The MTL approach was evaluated with an RCT involving 106 patients with depression, who were randomized in a 2:1 ratio to receive the integrated intervention. Our proposed MTL model outperformed both single-task models and existing multi-task models in predictive performance. Our approach represents a promising step in exploiting RCTs to develop predictive models for precision medicine.

Predicting mental disorders with wearables among a large cohort: Depressive and anxiety disorders are among the most prevalent mental disorders and are usually interconnected. We explored detecting those mental disorders with wearables in a large public dataset consisting of

more than 11,000 participants. We proposed a novel deep model that combines a transformer encoder and convolutional neural network to directly learn from the raw daily activity time-series data from the wearables. Our method achieved an area Under the Receiver Operating Characteristic curve (AUROC) of 0.717 (S.D. 0.009), demonstrating the feasibility of utilizing wearables to assist in diagnoses of mental health disorders.

Chapter 1

Introduction

1.1 Wearables in Healthcare Overview

Wearables, also known as "wearable devices", are generally referred to any miniaturized electronic devices that can be easily attached to and detached from the human body, or incorporated into clothing or other body-worn accessories [192]. Until now, there are a variety of off-the-shelf wearables, including smart earphones, smart rings and smartwatches, which have been pervasively adopted in fitness tracking, fashion style and entertainment. Nowadays, wearables also become appealing in healthcare. The growing awareness of healthcare and the aging society stimulate a high demand for medical services, but limited hospital resources hinder the access to professional healthcare for some people. Wearables can potentially fill this gap, via longitudinal remote monitoring outside of the clinical.

It is reported that the global market for healthcare wearables was USD 24.57 billion and will increase to more than 130 billion by 2026 with an annual growth rate of 24.7% [213]. The application of wearables is expected to become standard clinical practice. Emerging

wearable-based analytic platforms and artificial intelligence are facilitating automated health event prediction, prevention and intervention, extending professional healthcare from hospital intensive care unit (ICU) to resource-limited field settings in a patient-centric way. Figure 1.1 illustrates the landscape of wearables in healthcare.

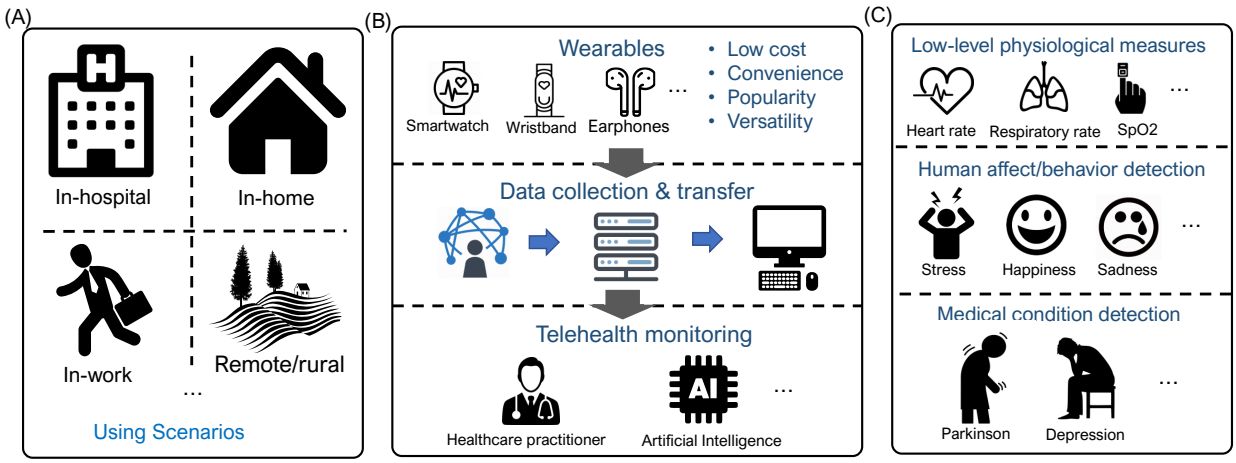


Figure 1.1: The wearable landscape.
(A) Various using scenarios; (B) Wearable data workflow; (C) Multiple levels of smart sensing in healthcare

The versatile onboard sensors from wearables enable endless possibilities for different health applications. The sensors majorly fall into two categories: mechanical, and physiological. Inertial measurement units (IMUs) are the most common mechanical sensor that can capture body motions. IMUs have been utilized to track gait [137], detect falls [71], and behavior patterns [102]. Physiological sensors, on the other hand, are aimed at directly measuring biological signals including vital signs. Photoplethysmography (PPG) is a common physiological sensor used to detect heart rate or oxygen saturation via changes in light absorbance through thin tissues. Some heart conditions (e.g., atrial fibrillation) can be detected with PPG sensors [181].

Usually, data collected from wearables are processed locally and/or transmitted to cloud storage for future analysis by healthcare practitioners and/or artificial intelligence. Processing the data locally ensures a better real-time performance, but it is only suitable for some tasks that do not require many computational resources. Offloading the data analysis to the cloud is beneficial to some complex and long-term disease monitoring. On cloud, we can adopt more advanced and sophisticated algorithms with access to much longer historical data, regardless of the resource limitations in wearable hardware.

Nonetheless, the role of wearables in healthcare is still premature. There remain challenges in unleashing the full power of wearables in healthcare. First, sensor readings from the wearables are vulnerable to motion and noise artifacts, which are exaggerated on commercial off-the-shelf wearables in ambulatory settings. Noise-contaminated signals could hinder producing accurate measurements, degrading the credibility of the wearables. To this end, a robust pipeline is needed to eliminate the adverse impacts of the noisy signals. Second, modern wearables are limited to some specific measurements, such as physical activities, heart rate, and oxygen saturation (i.e., SpO2). Although those measurements are reported to be associated with some clinical conditions [106, 192, 209], other important measurements (e.g., respiratory rate) are generally missing on the commercial wearables. We need to generate more clinically meaningful measurements from raw signals provided by the wearable sensors, as more measurements allow us to draw a more comprehensive picture for each patient. Finally, there exist gaps between the massive wearable data and the clinical insights. One general way is to build machine learning (ML) models with wearable data to predict personalized medical conditions. Those personalized estimations can then support clinicians to arrange treatments in a patient-centric way. For example, if the model predicts the patient has a high chance of having a bad outcome, the clinicians may prescribe an intervention treatment for the patient. However, establishing an accurate ML model with wearable data

is not straightforward. The fine-grained and noise-contaminated time series could easily make ML models overfitting. And sometimes, the models may also suffer from a small size of training samples.

1.2 Thesis Contributions

To address the above challenges and advance the application of wearables in healthcare, this dissertation investigates wearable applications from low-level wearable sensor measurements to high-level predictions of mental health disorders through the following contributions:

1.2.1 Enable Robust Respiratory Rate Measurements On Commercial Smartwatches

Modern smartwatches are equipped with Photoplethysmography (PPG) sensors to measure heart rate. However, the ability to accurately measure respiratory rate (RR) with motions is generally missing. Respiratory rate (RR) is a physiological signal that is vital for many health and clinical applications. In an ambulatory setting, the sensor readings from a wrist-worn smartwatch are pretty noisy, as wrist movements are inevitable. We presented the *RespWatch*, a wearable sensing system for robust RR monitoring on smartwatches with Photoplethysmography (PPG). We designed two novel RR estimators based on signal processing and deep learning. The signal processing estimator achieved high accuracy and efficiency in the presence of moderate noise. In comparison, the deep learning estimator, based on a convolutional neural network (CNN), was more robust against noise artifacts at a higher processing cost. To exploit their complementary strengths, we further developed a hybrid estimator that dynamically switches between the signal processing and deep learning estimators based on a new Estimation Quality Index (EQI). We evaluated and compared

these approaches on a dataset collected from 30 participants. The hybrid estimator achieved the lowest overall mean absolute error, balancing robustness and efficiency. Furthermore, we implemented RespWatch on commercial Wear OS smartwatches. The empirical evaluation demonstrated the feasibility and efficiency of RespWatch for RR monitoring on smartwatch platforms.

1.2.2 Predicting Objective and Subjective Stress With A Commercial Smartwatch

Built upon the sensor data processing pipeline from RespWatch, we explored stress detection on commercial smartwatches. Compared to respiration, stress is a more complicated physiological or psychological response, which can be categorized as objective stress or subjective stress [48]. The objective stress is defined as the biological reaction to a stressful exposure that manifests with biological reactions [177], whereas the subjective stress is defined as a subjective feeling of "being stressed". In this work, we described a methodological approach (a) to compare the prediction performance of models developed using objective markers of stress and subjective markers of stress; and (b) to develop personalized stress models by accounting for inter-individual differences. The objective stress markers were derived from a series of stressor tasks (e.g., a public speaking task or solving math problems) [151], and the subjective stress markers were derived from participants' subjective responses to self-reports (e.g., I am stressed). Towards this end, we conducted a laboratory-based study with 32 healthy volunteers. Our performance of the objective stress models with an instrumented commercial smartwatch was comparable to state-of-the-art models from other laboratory-based studies that require more sophisticated equipment. However, the generic subjective stress models had a lower performance compared to objective stress models. To enhance the subjective

stress prediction, we proposed the personalized subjective stress models accounting for inter-individual differences via an adaptive threshold. Unlike traditional personalized models, our approach does not need to build a machine learning model for each user. Results demonstrated our personalized subjective stress model has significant performance improvements over the generic subjective model.

1.2.3 Multi-Task Learning for Randomized Controlled Trials with Wearables

A randomized controlled trial (RCT) is commonly regarded as the ultimate tool to validate treatment by comparing patients' outcomes in an intervention group and a control group [55]. Previous statistical methods used for RCT analysis lack the ability to assess the treatment response at an individual level. In this paper, we exploit machine learning models in conjunction with RCTs for personalized predictions of a depression treatment in which patients were monitored with wearable data. We formulated the predictions in different groups from an RCT as a multi-task learning (MTL) problem, and proposed a novel MTL model specifically designed for the RCT. Instead of training separate models for the intervention and control groups, our MTL model can be trained on the combined groups of patients, effectively enlarging the training dataset. We devised a hierarchical model architecture to aggregate data from different sources and different stages of the trial, which allows the MTL model to exploit the commonalities and capture the differences between two groups in the RCT. We evaluated the MTL approach in the RCT involving 106 patients with depression, who were randomized in a 2:1 ratio to receive the integrated intervention. Our proposed MTL model outperformed both single-task models and existing multi-task models in predictive performance. Our approach represents a promising step in exploiting RCTs to develop predictive models for precision medicine.

1.2.4 Predicting Mental Disorders with Wearables: A Large Cohort Study

Depression and anxiety are among the most prevalent mental disorders, and they are usually interconnected [40]. Although those two mental disorders have drawn increasing attention due to their tremendous negative impacts on working ability and job performance [117], over 50% of patients are not recognized or adequately treated. Recent literature has shown the potential of using wearables for expediting the detection of mental health disorders [67], as physical activities are reported to be related to some mental health disorders [9]. However, most prior studies [151, 209] focused on a single group of people with limited sample size. The feasibility of using wearables to detect mental disorders in the general public remains questionable. We explored detecting depression and anxiety disorders with commercial wearable activity trackers in a large public dataset consisting of more than 11,000 people. The dataset has a wide spectrum of age, race, ethnicity, and education levels. We proposed a novel deep model that combines a transformer encoder and convolutional neural network, which can directly learn from the raw wearable data. Our method can achieve an area Under the Receiver Operating Characteristic curve (AUROC) of 0.717 (S.D. 0.009), affording new opportunities in using wearables to assist in the diagnosis of mental health disorders.

1.3 Dissertation Organization

My dissertation is structured as follows. Chapter 2 first presents *RespWatch*, which targets robust measurements of respiratory rate on the commercial smartwatches. In this work, we introduced a hybrid approach that exploits the complementary advantages of signal processing and deep learning techniques to handle the noisy raw signal data from the sensor. Then in Chapter 3, we introduced two categories of stress detection models on the commercial

smartwatches. The model pipeline was built upon noise elimination techniques in Chapter 2. Chapter 4 demonstrates the application of wearables in a randomized controlled trial to predict the patients' outcome in two arms (i.e., control and intervention). Chapter 5 presents the application of wearables in the detection of depressive and anxiety disorders among a large cohort. Finally, Chapter 6 concludes this dissertation.

Chapter 2

Robust Measurements of Respiratory Rate on Smartwatches

In this chapter, we present *RespWatch*, which targets at using smartwatches to measure a low-level physiological signals (i.e., respiratory rate). We proposed a robust signal processing pipeline to handle the noisy raw sensor data, which is also the foundation of our work in Chapter 3.

2.1 Introduction

Respiratory rate (RR) is an important physiological variable associated with serious health conditions such as cardiopulmonary arrest [43]. In addition to the clinical applications, RR is important for ascertaining driving safety [104, 216], assessing sleep quality [13], monitoring stress [65] and even detecting opioid overdose [125]. However, unobtrusive monitoring of RR outside of laboratory and hospital settings is difficult. Traditional approaches for RR measurements rely on the use of the specialized equipment, e.g. capnography system and

nasal/oral pressure transducers[22]. These approaches are not suitable for "free-living" or long-term measurement outside controlled clinical environments. Robust RR measurements with a popular commercial device can renovate the approaches to the real-time detection and long-term monitoring of respiration-related health conditions.

In this paper, we address the problem of robust RR monitoring using photoplethysmography (PPG) sensors on commercial smartwatches. The adoption of wearable devices, and smartwatches in particular, has increased exponentially over the past decade [20]. PPG sensors have been commonly embedded in smartwatches to measure heart rate and detect various health conditions, such as atrial fibrillation [181] and sleep apnea [86]. And smartwatches have the potential to enable unobtrusive longitudinal RR monitoring outside clinical environments with the PPG sensor.

However, RR monitoring on smartwatches with PPG faces several challenges. First, many previous studies [85, 109, 149] focused on PPG sensors for measuring light signals transmitted through fingertips, whereas smartwatch PPG sensors measure signals reflected from the wrist, which degrades signal quality and introduces noise artifacts [160]. As such, it is essential to develop robust approaches to extract RR from noisy PPG signals [162], and to investigate the feasibility of reliable RR measurements on off-the-shelf smartwatches. Second, previous research on RR monitoring with PPG usually targeted use cases with minimum or no motion (e.g., a patient wearing a pulse oximeter in an Intensive Care Unit (ICU) bed) [85, 150, 200], whereas we aim for RR monitoring in the presence of some user motions and noise artifacts. It is inevitable to the motions with a wrist-worn smartwatch in the unconstrained settings. Therefore, smartwatch-based RR measuring system must be consistently robust for the longitudinal monitoring. Finally, smartwatches have limited computational resources and power. For any real-time and long-term RR monitoring system running on the smartwatch,

data processing pipelines and algorithms should be highly efficient and capable of continuous execution on the resource-constrained platform.

Towards this end, we present *RespWatch*, a wearable sensing system for robust RR measurements with built-in PPG sensors on commercial smartwatches. RespWatch provides end-to-end processing pipelines from the raw PPG signals to RR measurements that can maintain high accuracy in the presence of some noise and motion artifacts. We explore and compare both signal processing and deep learning approaches, and develop a hybrid approach to combine their complementary strengths. Furthermore, RespWatch is capable to run completely on commercial smartwatches which allows for non-obtrusive RR monitoring. Specifically, the main contributions of this research are as follows:

- A *signal processing estimator* with fine-grained elimination of noise artifacts, which achieves efficiency and accuracy in the presence of moderate noise artifacts;
- A *deep learning estimator* for extracting RR from noisy PPG signals, which exhibits robustness in the presence of increasing noise artifacts;
- A *hybrid approach* which dynamically switches between signal processing and deep learning based on a novel Estimation Quality Index (EQI), achieving both robustness and efficiency;
- A comparative evaluation of the RR estimation approaches on a dataset including 30 participants of various activities, which demonstrates the complementary strengths of the signal processing and deep learning estimators and the advantage of combining both approaches in the hybrid estimator;

- Implementation and experimentation of RespWatch on commercial Wear OS smartwatches, which demonstrates the feasibility and efficiency of RR monitoring on smartwatch platforms.

2.2 Related Work and Background

2.2.1 Non-contact RR Measurement

Recently, non-contact sensing approaches have been developed for measuring RR. Techniques based on radio frequency (RF) detect respiration based on changes in RF signals caused by inhalation and exhalation motions. RR has been estimated using Frequency Modulated Carrier Waves (FMCW) [210] and Doppler radar [219]. WiFi signals have also been adopted to estimate RR based on the received signal strength (RSS) [144] and channel state information (CSI) [211]. Other non-contact sensing techniques for RR measurement exploit energy spectrum density (ESD) of acoustic signals [216] and ground movement from geophones [83]. As these non-contact approaches rely on external devices in the environment, they are constrained to instrumented environments and cannot provide monitoring when users leave such environments.

2.2.2 IMU-based RR Measurement

Smartwatches provide a portable platform with built-in sensors that can be utilized for unobtrusive sensing. Previous research [54, 64, 107, 197] on RR measurement with smartwatches exploited the inertial measurement unit (IMU) to capture subtle motions owing to respiration. However, this micro-motion is easily overwhelmed by motion artifacts [107] during normal activities. Hence, IMU-based RR monitoring is usually limited to constrained settings with

minimum motion. For instance, Sun et al. [197] designed a total variation filter to extract respiratory signals from accelerometer data captured by smartwatch during sleep. Similarly, Hao et al. [54] developed the MindfulWatch to monitor respiratory during meditation, using a similar filtering approach. To extend the RR measurements in daily living activities, Liaqat et al. [107] proposed to identify accurate sensor readings with respiration information using a machine learning model, and extract RR only from those accurate sensor readings. However, since the micro-motions associated with respiration could be of the same order of magnitude as the sensor noise [54] and several orders of magnitude lower than other body motions, the signal-to-noise ratio (SNR) can often drop below the threshold for valid measurements.

2.2.3 PPG-based RR Measurement

PPG is an optical sensing technology that detects pulsatile blood volume changes in tissues [165]. Compared to IMU, PPG sensor readings are less vulnerable to motion artifacts, as PPG measures the optical changes that are not directly impacted by the motions. As illustrated in Figure 2.1, a PPG sensor consists of a light-emitting diode (LED) to illuminate the tissue and a photodiode (PD) to measure the light transmitted through or reflected by the tissue. Transmission-mode PPG is commonly used in fingertip pulse oximeters, whereas reflectance-mode PPG is usually used on wrist or forehead for heart rate monitoring. The mode and placement of the PPG sensor has impacts on its sensing accuracy and waveform shape [160].

RR measurement is based on the fact that the PPG waveform is modulated by the respiration process. As illustrated in Figure 2.2, the PPG waveform contains three types of respiratory-induced variations caused by amplitude, intensity and frequency modulation [85, 119].

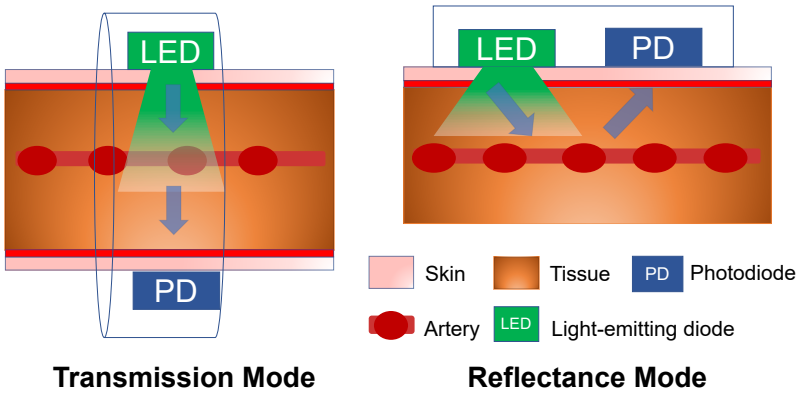


Figure 2.1: Two modes of the PPG sensor.

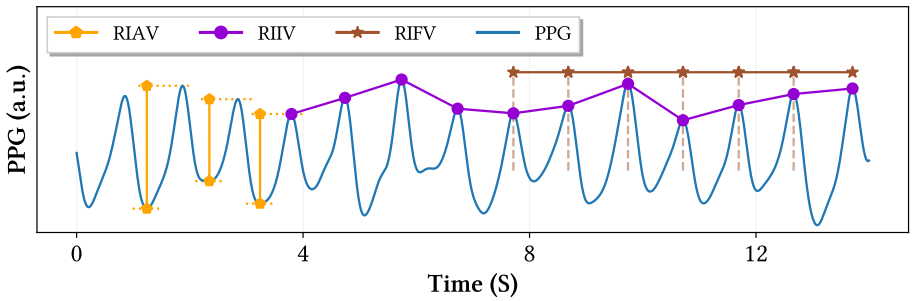


Figure 2.2: PPG waveform and respiratory-induced variations. RIAV: respiratory-induced amplitude variation; RIIV: respiratory-induced intensity variation; RIFV: respiratory-induced frequency variation.

- Amplitude modulation leads to *respiratory-induced amplitude variation (RIAV)*, which is related to changes in peripheral pulse strength [85]. RIAV is reflected in the different amplitudes of the peaks and corresponding valley for each pulse in the PPG waveform, and can be extracted as a time-series of the vertical distances from the peak to the valley for each pulse.

- Intensity modulation leads to *respiratory-induced intensity variation (RIIV)*, which is related to the intrathoracic pressure variation [85]. RIIV is reflected in a baseline wander [119] in the PPG waveform, and can be extracted as a time-series of the peak heights.

- Frequency modulation leads to *respiratory-induced frequency variation (RIFV)*, which is related to an autonomic response to respiration. RIFV, also referred as respiratory sinus arrhythmia (RSA) [85], is reflected in different inter-beat intervals, and can be extracted as a time-series of the horizontal distances between the successive peaks in the PPG waveform.

RR can be estimated in two general steps [21]: (1) extracting the respiratory variation signals, and (2) estimating of RR from the variation signals. Karlen et al. [85] used the Incremental-Merge Segmentation method to detect artifacts and extract the three respiratory-induced variations (RIIV, RIAV, RIFV). RR can then be obtained from the variations by Fast Fourier Transform (FFT) with smart fusion. Pimentel et al. [150] improved the reliability of RR measurements with multiple autoregressive (AR) models for determining the dominant respiratory frequency in the three variations. Compared to the fusion method from [85], the AR models can retain more data windows. The aforementioned studies on PPG-based RR measurements [21, 85, 109, 150] focused on fingertip sensors in clinical settings or during sleep with limited motion. Since the signal admission control in those studies discards entire sampling windows affected by noise artifacts, the approaches may lead to low data yield in the presence of user activities. How to robustly distill respiratory information from the raw PPG signal remains challenging especially in the presence of noise.

Video-based PPG has also been explored to measure RR with smartphone cameras[167]. Due to its reliance on video taken by cameras, this approach is not suitable for long-term and non-obtrusive RR monitoring during daily activities.

Recent studies [80, 110, 200] applied similar signal processing approaches to measure RR with reflective PPG sensors. Jarchi et al. [80] and Longmore et al. [110] explored measuring RR at different body positions (including wrist) with reflective PPG sensors. They found that

upper-body positions (e.g., head and neck) produced the best respiration signals. Trimpop et al. [200] demonstrated a system on commercial smartwatches for RR monitoring during sleep and evaluated it on four users, but without revealing the details of the methodology. As those studies [80, 110, 200] adopted similar signal processing approaches to those developed for fingertip PPG, they did not address the more significant noise artifacts with user activities and the reflectance mode of PPG sensor. In contrast, we present novel signal processing techniques specifically designed to robustly estimate RR in the presence of noise artifacts and user activities. Moreover, we explore deep learning to further enhance the robustness of RR measurements against noise and motion artifacts, and integrate both approaches to balance robustness and efficiency of RR monitoring on commercial smartwatches.

2.2.4 Deep Learning on Smartwatch

Deep learning with wearable data has drawn great attentions in activity recognition [11, 157], Parkinson Disease monitoring [53, 202], atrial fibrillation detection [148, 181] and other mobile health applications [107, 216]. Ravichandran et al. [163] had proposed a dilated residual inception model to regress the respiration waveform from the PPG waveform. But their study was limited to the fingertip PPG signals collected in the intensive care unit (ICU), and cannot estimate the respiratory rate directly. Those application-driven studies have demonstrated that deep learning is capable to handle some sophisticated problems with the wearable data. However, further empirical evaluations of the deep learning models are required to test their capability of running on the wearable devices in real life.

2.3 Design of RespWatch

Towards RR measurements outside the clinic settings, our RR monitoring system shoots the following goals:

- **Accuracy.** The system should produce accurate RR measurements.
- **Robustness.** The system should maintain accuracy and data yield in the presence of noise artifacts.
- **Efficiency.** The system should have light-weight and efficient processing pipeline on smartwatches.

We exploited both signal processing and deep learning approaches to the RR estimations. In this section, we first design a signal processing estimator that achieves efficiency and accuracy in the presence of moderate noise artifacts. We then build a deep learning estimator that is more robust against increasing noise artifacts while incurring higher processing cost. Finally, we develop a hybrid estimator that balances robustness and efficiency by dynamically switching between the signal processing and deep learning estimators.

2.3.1 Signal Processing Estimator

The signal processing estimator employs digital signal processing techniques, which are training-free and allow for efficient processing on a commercial smartwatch. We designed a signal processing pipeline comprising three stages (see Figure 2.3).

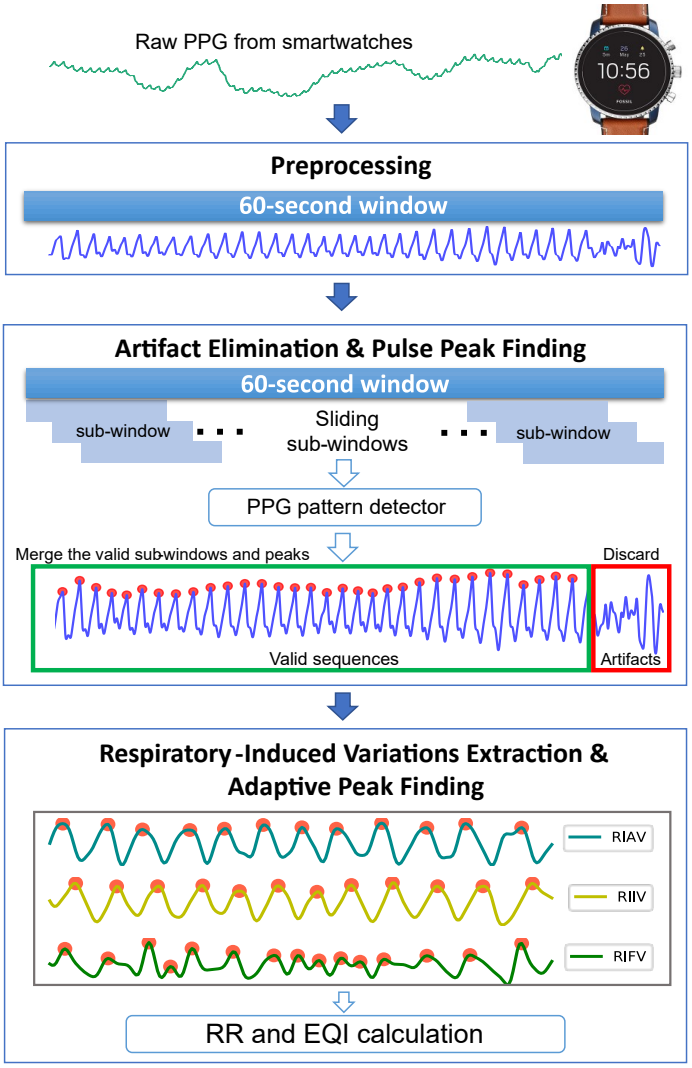


Figure 2.3: Architecture of the signal processing estimator in RespWatch

1) In the *preprocessing* stage, we use a bandpass filter to eliminate noise outside the cardiac and respiratory bands from the raw signals, and then divide the signal waveform into 60-second windows.

2) In the *artifact elimination and pulse peak finding* stage, we employ a fine-grained technique to remove data points corrupted by noise artifacts within the cardiac and respiratory bands. We then use a novel *PPG pattern detector* to find the pulse peaks in the remaining data points. To facilitate finding the pulse peak, the PPG pattern detector employs a novel

forward-backward highpass filter to remove respiratory band information while preserving the time-domain features. This allows the identified pulse peaks to be mapped to the original waveform with the respiratory information.

3) Finally, we extract the respiratory-induced variation signals, using the pulse peak positions found by the PPG pattern detector. Then, we mapped the variation signals to the RR estimations with an adaptive peak finding method. In addition, we introduce a novel *estimation quality index (EQI)* to assess the accuracy of our RR measurements, which enables the hybrid estimator to dynamically switch between signal processing and deep learning estimators, maintaining high accuracy and efficiency.

In the following we detail the design of each stage.

Preprocessing

The sampling rates of the PPG sensors vary across smartwatches and are usually higher than 100 Hz. In order to minimize the effect of the differences of sampling rates, we re-sample the collected PPG data at a fixed rate of 50 Hz based on data timestamps. The raw PPG waveform contains many noisy frequency components. A sixth-order Butterworth bandpass filter is first applied to remove the unwanted noisy components with cut-off frequency of 0.14Hz and 3Hz, only keeping the information from respiratory band to cardiac band [126]. The preprocessing does not remove noise artifacts within the ranges of respiratory and cardiac bands, which is handled in the next stage. The signal is then re-scaled to the range from -1 to 1. We divide the PPG data into 60-second windows for RR estimation in following stages. There is no overlap between the adjacent windows. The 60-second length has been used for RR studies in previous literature [85, 150].

Artifact Elimination and Pulse Peak finding

Noise artifacts are inevitable on smartwatches due to the wrist movement and poor contact between the sensor and skin. Previous research [21, 85, 109, 150] has often discarded any data window containing noise artifacts. Although this approach can help in avoiding noise artifacts, it leads to significant drop in RR measurement yield, especially during user activities. To support long-term RR monitoring in the presence of noise artifacts and improve data yield, we introduce a *sliding sub-window* technique to discard noise artifacts at a finer granularity while preserving the valid data samples in the same 60-second data window. The sliding sub-window has a size of 10 seconds and a step size of 2 seconds. Each 10-second sub-window from the 60-second window is passed through the PPG pattern detector to evaluate whether it is free from noise artifacts, and to identify valid pulse peaks simultaneously. The entire procedures for artifact elimination and pulse peak finding are summarized in Algorithm 1.

Algorithm 1: Artifact Elimination & Pulse Peak Finding

Data: 60-second preprocessed PPG waveform

Sliding sub-windows with size of 10s and step of 2s;
for *each sub-window* **do**
    `/* begin of PPG pattern detector */`
    2nd-order highpass **forward-backward filtering**;
    Re-scale the waveform with range of [-1,1];
    Find the *peaks* higher than 0;
    Calculate *heart rate*, *peak intervals*, *peak-to-valley distances*;
    `// PPG pattern matching`
    if *heart rate* $\leq$ *180* ***and*** *heart rate* $\geq$ *40* ***and*** *STD(peak intervals)<0.4s* ***and*** *STD(peak-to-valley distances)<0.4* **then**
        mark the sub-window as valid ;
    end
end
Merge the consecutive valid sub-windows into valid sequences;
Merge the peaks from the valid sequences into peak lists;

Result: Valid sequences and corresponding peak lists

In the PPG pattern detector, we first filter out the respiratory band information in the sub-window, as the respiratory band can impact the accuracy of finding the pulse peaks. A novel design of the PPG pattern detector is the adoption of a second-order forward-backward highpass Butterworth filter with cut-off frequency of 0.6 Hz. The forward-backward filter is a *zero-phase* filter in which the phase response slope is zero at all frequencies. It achieves the zero-phase response by filtering the input data twice, first in the forward direction and then in the reverse direction. Hence, the order of the filter is doubled, and the filter is non-causal due to the reverse filtering [190]. Since the processing of sub-windows is performed once we have the 60-second large window, there is no requirement of the causality of the filter.

The forward-backward filter is a key component of the PPG pattern detector. A significant benefit of the zero-phase filter is that it is able to preserve important time-domain features in the filtered signal. Specifically, the pulse peaks in the filtered waveform appear at the same positions as the pulse peaks in the unfiltered waveform in the time domain [33]. This allows us to directly map the pulse peaks found in the filtered waveform back to the unfiltered waveform. Consequently, we can find the pulse peaks in the unfiltered waveform in the 60-second window when iterating through the sub-windows containing the filtered signals. The pulse peaks in the unfiltered waveform will be used to extract the respiratory-induced variation time-series in the next stage.

Figure 2.4 shows the advantage of the forward-backward filter with the real PPG data collected as part of our study. The blue curve shows the unfiltered waveform containing the respiratory band components. It has a large respiratory-induced baseline wander. Additionally, the pulses are not clearly distinguishable, making it difficult to find pulse peaks using standard peak finding methods[84, 183]. The green curve shows the waveform after it is processed using our forward-backward filter. Clearly, the peaks are distinguishable making it easier to check whether the waveform contains valid PPG patterns with a pattern matching method,

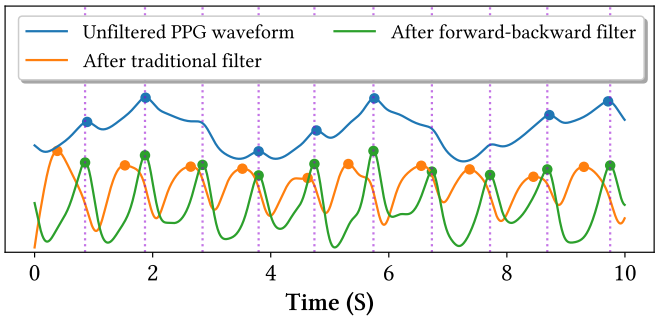
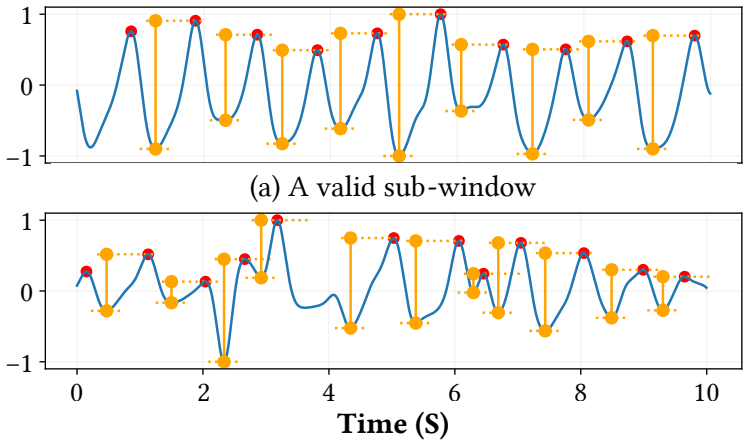


Figure 2.4: Comparison of preprocessed PPG waveform with traditional Filter and forward-backward filter.

and identify each peak by detecting local maxima above a certain threshold. It is important to note that these pulse peaks can be mapped back to the unfiltered curve (see the vertical dashed lines in Figure 2.4). In contrast, the orange curve is the filtered waveform after being processed by a traditional fourth-order Butterworth filter with the same cut-off frequency. We can observe variable time shifts in the peak positions after the traditional filter, making it impossible to map the pulse peaks back to the blue curve. As a result, the forward-backward filter not only removes the respiratory band to facilitate PPG pattern matching and peak finding, but also allows the mapping of pulse peaks back to the unfiltered PPG waveform containing the raw respiratory information.

After forward-backward filtering, we re-scale the signal to a range of $[-1, 1]$, and identify all the peaks whose amplitude are higher than 0. Then, we implement our PPG pattern matching method derived from [135] to detect whether the PPG signal is valid using three rules: (1) extracted heart rate based on the peaks should be within 40 and 180 bpm; (2) the standard deviation of the peak intervals should be less than 0.4s; (3) the standard deviation of the *peak-to-valley* distances should be less than 0.4, where the peak-to-valley distance is the vertical distance from the peak to its previous valley. We find the valleys via the local minimum between two adjacent peaks. Only those sub-windows satisfying all three rules are marked as valid.



(a) A valid sub-window

(b) An invalid sub-window with artifacts. (The standard deviation of *peak-to-valley* distances is larger than 0.4.)

Figure 2.5: Examples of PPG pattern matching.

Figure 2.5 shows real examples of a valid sub-window and an invalid sub-window with artifacts identified by the PPG pattern detector. The valid sub-window contains pulses satisfying the aforementioned rules, whereas the invalid sub-window has the standard deviation of the *peak-to-valley* distances larger than 0.4, not satisfying the third rule.

Once we have iterated through all the sub-windows with the PPG pattern detector, consecutive valid sub-windows are merged into larger valid sequences, and the peaks in the valid sub-windows are also merged into longer lists of peaks (see Figure 2.3). Here, the valid sequences mark the start and end points of a preprocessed PPG waveform free of noise artifacts, and the peak lists contain timestamps of the pulse peaks in the corresponding valid sequences. A 60-second preprocessed PPG waveform can have multiple valid sequences and valid peaks lists, if the invalid sub-windows appear in the middle of the 60-second period.

Respiratory-induced Variations Extraction and Adaptive Peak Finding

As described in Section 2.2.3, the respiratory-induced amplitude variation (RIAV) is the time-series of the vertical distances from the peak to the valley for each pulse. The respiratory-induced intensity variation (RIIV) is the time-series of the height of each pulse peak. The respiratory-induced frequency variation (RIFV) is the time-series of the horizontal distances between successive pulse peaks in the time domain. All the three variation time-series are closely related to pulse peaks that need to be extracted from the preprocessed PPG waveform containing the respiratory band information. We directly adopt the peak lists from the last stage, and map the pulse peaks from the filtered waveform to the preprocessed PPG waveform. The valleys in the preprocessed waveform are then obtained by finding the local minimum between two adjacent peaks. The adoption of peak lists from the second stage improves the accuracy of extracting the three time series, because we avoid directly finding pulse peaks in the preprocessed PPG waveform in which the respiratory band information can degrade the accuracy of pulse peak finding. And it also saves us from the pulse peak finding twice for the PPG pattern matching and respiration signal extraction, making the system more energy-efficient.

The time-series of RIAV, RIIV and RIFV are not equally sampled in time domain, so we re-sample the three at $f_s = 5Hz$ with linear interpolation. We also employ a bandpass filter to keep only the respiratory band (0.14-0.9Hz) information. To robustly detect all the respiratory peaks in the RIAV, RIIV and RIFV, we apply an adaptive peak finding method derived from [84]. The method starts with initial thresholds for the distance between two adjacent peaks. We find all the peaks whose amplitude are higher than 0, and calculate the horizontal distance from the current peak to the last peak. If the distance is below the lower threshold, the current peak will be discarded and the lower threshold decreases. If the

distance is beyond the higher threshold, a virtual "peak" will be inserted in the middle of the current peak and the last peak, and the higher threshold increases. The initial thresholds and the adjusting rates are set based on [126]. Adaptive peak finding method can handle the cases in which the artifacts cause a spurious peak or obliterate a possible peak in the signals, based on the assumption that the RR is constant within a short period of time.

After we get the respiratory peaks in the three variations, the respiratory rate (RR) is calculated for each valid sequence:

$$RR_{RIXV,i} = \frac{60}{MEAN(peak\_intervals_{(i)})/f_s} \tag{2.1}$$

where $RIXV$ is one of the three respiratory-induced variations, $MEAN(\cdot)$ is the average value of $\cdot$, f_s is the sampling rate, i is the index of the valid sequences, and $peak\_intervals_{(i)}$ is the respiratory peak intervals detected by the adaptive peak finding method for the i^{th} valid sequence. The final RR measurement $RespWatch_{RIXV}$ is the length-weighted average of $RR_{RIXV,i}$:

$$RespWatch_{RIXV} = \frac{\sum_i RR_{RIXV,i} \cdot seq\_length_{(i)}}{\sum_i seq\_length_{(i)}} \tag{2.2}$$

Estimation Quality Index (EQI)

Furthermore, we introduce an *estimation quality index (EQI)* as a novel metric to estimate how accurate our RR measurement is. EQI is based on the two intuitions: (1) the respiration is rhythmic, so the standard deviation of the respiration peak intervals should be small, and (2) RR measurement is more accurate on the longer sequence. Specifically, the EQI of each valid sequence is formulated as:

$$EQI_{RIXV,i} = \alpha \cdot \frac{STD(peak\_intervals_{(i)})}{seq\_length_{(i)}} \tag{2.3}$$

where α is a fixed scaling factor, $STD(\cdot)$ is the standard deviation of $\cdot$, $seq\_length_{(i)}$ is the length of the i^{th} valid sequence. The final EQI_{RIXV} is the sum of $EQI_{RIXV,i}$ for each valid sequence:

$$EQI_{RIXV} = \sum_{i} EQI_{RIXV,i} \tag{2.4}$$

EQI offers several important advantages. First, most prior studies only focus on the RR estimation without providing an accuracy estimation. Lack of confidence of the measurement could lead to wrong decision in some practical cases. For example, the inaccurate high RR measure could give a false alarm of respiration conditions. Second, although past studies focused on motion as the main factor influencing PPG-based sensing accuracy [162, 181], other factors (e.g., light conditions and sweat) may also affect accuracy. EQI therefore captures noise artifact in a more comprehensive manner than motion artifacts alone. Finally, as EQI utilizes only the characteristics of the RR estimation process itself, it does not require external inputs (e.g., motion intensity, light, and sweat).

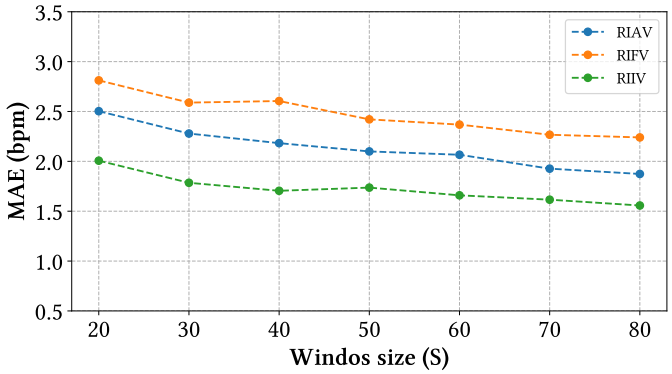


Figure 2.6: MAE vs. window sizes. All the data windows are free from artifacts

To assess our assumption that the accuracy of RR measurement improves with larger data windows, Figure 2.6 shows the mean absolute error (MAE) of RR measurements with different window sizes. From data collected from our user study (see Section 2.4), we randomly sampled 100 data windows at each window size, and all the sampled data windows were free from artifacts. We can observe that the MAE decreases with larger window size, which supports our assumption.

2.3.2 Deep Learning Estimator

This section presents the deep learning estimator for RR measurement. Our work was inspired by recent success of deep learning in processing smartwatch data [107, 181]. Particularly, Shen et al. [181] showed that a convolutional neural network (CNN) model with residuals was robust in the presence of motion artifacts for detecting atrial fibrillation with the smartwatch PPG. Building upon this, we designed the deep learning estimator with a CNN model. After some basic preprocessing steps, our CNN model can directly output the estimation of RR using the PPG waveform. To the best of our knowledge, our deep learning estimator is the first deep neural network to estimate RR with wrist PPG on smartwatches. The high-level architecture of the deep learning estimator is illustrated in Figure 2.7.

Preprocessing

Although the CNN model can directly learn from raw signals, preprocessing is still needed to account for issues associated with the PPG signals. Specifically, raw PPG signals exhibit different ranges and amplitudes under different conditions, and the noise outside the respiratory and cardiac bands can lead to the overfitting of the CNN model. To standardize data and reduce noise, we re-sampled the PPG signal at 50Hz, applied the same bandpass filter used

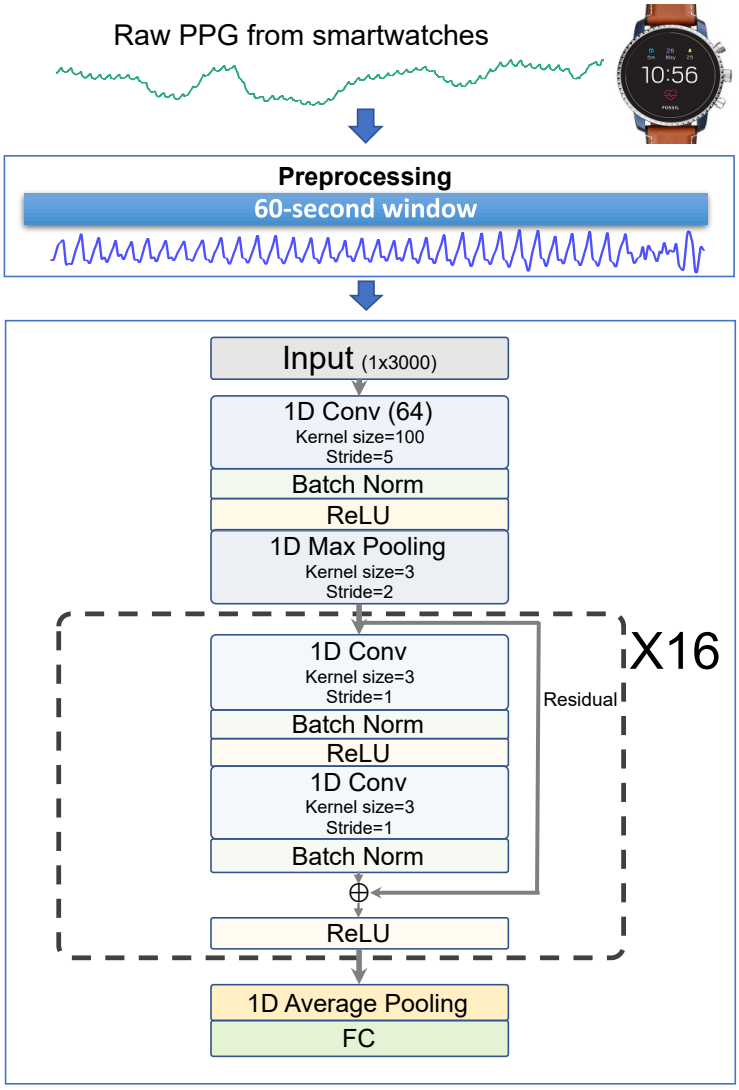


Figure 2.7: Architecture of the deep learning estimator in RespWatch

in the signal preprocessing estimator, and normalized the signals to a zero mean and a unit variance.

CNN Model

As shown in Figure 2.7, we developed our deep learning approach based on the residual neural network. Unlike previous classification tasks with PPG [181], our aim is to output the RR estimation with continuous values. As the PPG sensor on smartwatches contains only one channel, we adopted 1D convolutional layers across the network.

An initial 1D convolutional layer with kernel size of 100 is adopted to down-sample the input and reduce the computation complexity. Then, 16 basic blocks sharing the same topology with residuals bypass and 1D convolutions are applied. Each basic block contains 2 convolutions and a shortcut connection. The shortcut connections can optimize the training by allowing information to propagate in deep neural networks [59] and make the optimization process tractable. Batch normalization (Batch Norm) and a rectified linear unit (ReLU) activation layer are also employed after each convolutional layer. The 16 basic blocks are grouped into 4 stages consisting of 3, 4, 6 and 3 blocks, and the number of output channels for each stage is 64, 128, 256 and 512, respectively. The spatial map of the signal is down-sampled while the channels are incremented stage by stage. After the last stage of the basic blocks, we append a 1D average pooling layer and a fully connected layer. The fully connected layer performs the regression tasks of the final RR estimation. We employed the mean squared loss and the stochastic gradient descent optimizer with the momentum. During training, we ensured there is no overlap between the training and testing signal. All the convolutional layers were initialized with *He* initialization [60], and batch normalization layers were initialized with weight of 1, bias of 0.

2.3.3 Hybrid Estimator

A key finding in our experimental results (see Sections 2.5 and 2.6) is that the signal processing and deep learning estimators have complementary strengths in efficiency and robustness, respectively. Specifically, the signal processing estimator achieves higher accuracy in the presence of moderate noise artifact. It also incurs lower processing cost on smartwatch platforms. In contrast, the deep learning estimator exhibits more robustness against increasing noise artifact. To maintain accuracy, robustness, and efficiency under varying noise artifact, we developed the hybrid estimator to combine the strengths of both the signal processing and deep learning estimators. Under increasing noise artifact, the hybrid estimator automatically switches from signal processing to deep learning to take advantage of its higher level of robustness. Conversely, the hybrid estimator switches back to signal processing when noise artifact diminishes to benefit from its higher efficiency and accuracy.

The key to the design hybrid estimator is to identify the metric used to make the switching decision. We explored motion intensity and EQI as two alternative metrics used to choose between the two estimators. Motion intensity is defined as the standard deviation of the magnitude of the tri-axial acceleration in a 60-second window [181]. It can be obtained from the IMU sensor in smartwatches. In comparison, as defined in Section 12, EQI characterizes the estimation quality that may be influenced by noise artifact in general, which may include both motion and other sources of noise (e.g., poor sensor contact).

For either motion intensity or EQI, we applied a grid search to find the best switch threshold that leads to the lowest mean absolute error (MAE) in RR measurements. We found experimentally (see evaluation in Section 2.5) that EQI outperforms motion intensity in accuracy and efficiency. In addition, EQI is derived from the PPG signal itself, which does not rely on external sensors such as the IMU.

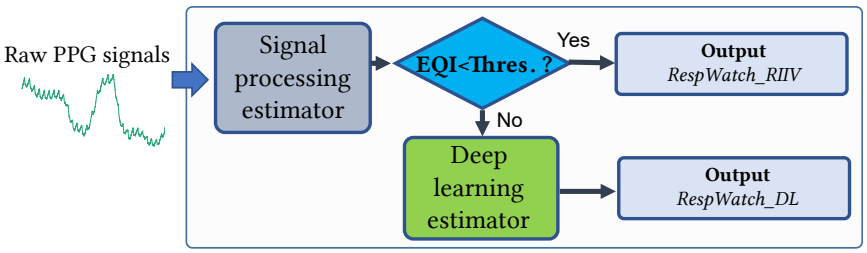


Figure 2.8: Architecture of hybrid estimator. *RespWatch_RIIV* is the output from signal processing estimator with RIIV; *RespWatch_DL* is the output from deep learning estimator.

The EQI-based hybrid estimator is illustrated in Figure 2.8. The EQI is first generated from the signal processing estimator. Since the signal processing method is highly efficient, the execution of the signal processing estimator incurs minor overhead for the whole system. After we get the RR measurement and EQI from signal processing estimator, if the EQI is below the switching threshold, the hybrid estimator directly output RR measurement from signal processing estimator. Otherwise, it invokes the deep learning estimator to produce the RR measurement.

2.4 User Study

We collected PPG data through a user study involving 32 healthy volunteers. The data collected in this study was primarily used to evaluate the accuracy of the RR estimations (see Section 2.5). The run-time efficiency of the estimators was empirically evaluated on smartwatches in Section 2.6.

2.4.1 Devices

We instrumented mainstream smartwatches, Fossil Gen4 Explorist, to collect raw PPG signals used to evaluate the RR estimations from RespWatch. The ground truth was obtained with Vernier Respiratory Go Direct Respiration Belt, which was used in the previous respiration

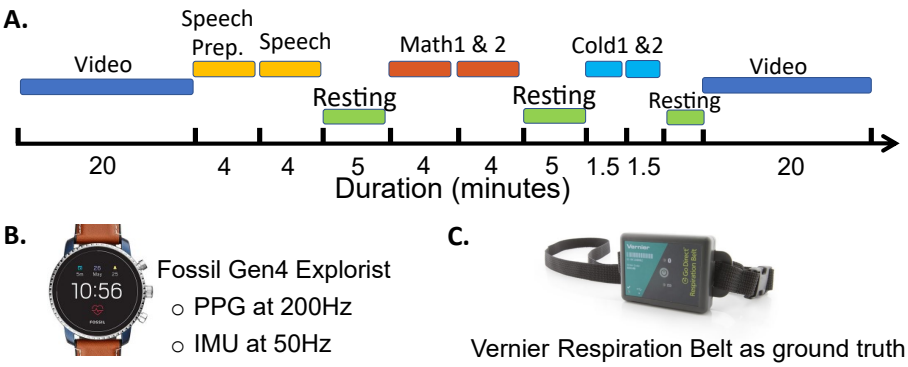


Figure 2.9: (A). Sequence of activities of the collecting procedure. (B). Fossil Gen4 Explorist smartwatch instrumented for this study. (C). Vernier Respiratory Go Direct Respiration Belt as ground truth.

studies [36, 123]. We also collected acceleration data from the IMU on the watch to measure motion intensity during the study. During data collection, each participant was asked to wear the belt over their chest and the smartwatch on their non-dominant hand. A custom application was installed on the smartwatches to record the data from the PPG sensor at 200Hz and IMU at 50Hz. The data were initially stored locally on the smartwatches, and then uploaded to a secure server.

2.4.2 Study Protocol

32 healthy volunteers were recruited through flyers posted across the campus at Washington University in St. Louis. All the participants met the inclusion criteria (between 18 and 69 years of age, with no heart disease, not pregnant at the time of recruitment, and not having an implanted pacemaker). At the end of the study, a compensation of a $25 was provided. The institutional review board (IRB) of Washington University in St. Louis approved this study, and written consents were obtained from all participants (IRB#2019-04150). The data was collected in various scenarios, including (1) watching a video, (2) preparing and delivering a speech, (3) doing mathematical tasks on computers, and (4) holding a cold object

for an extended period. All the activities involved motions to same degree. The timeline of the data collecting procedure and the devices is shown in Figure 2.9.

Two participants' data were lost due to issues during data upload. As a consequence, only 30 participants' data were used in the analysis and evaluation. Additionally, we exclude the data when the ground truth is not available, i.e., the Respiration Belt failed to acquire valid RR measurements. This occurred during some segments when participants were delivering the speech, as speaking caused unreliable RR measurements [65].

2.4.3 Impacts of Activities

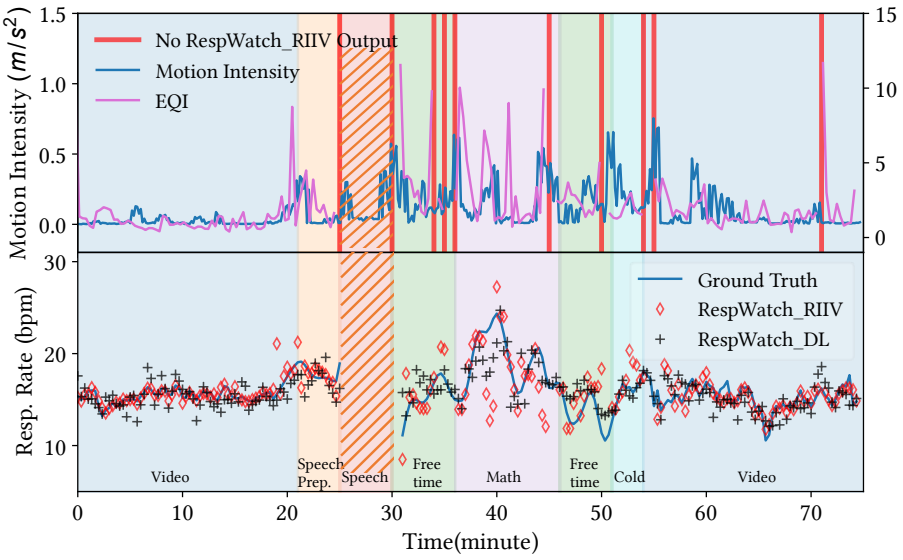


Figure 2.10: RR measurements, motion intensity, and EQI of one user participating in the study.

To show the impacts of noise or activity on RR measurements, we plot the RR time series produced by the RespWatch estimators for one user over the entire session, as shown in Figure 2.10 (excluding the speech activity as mentioned in the last subsection). The top

graph shows the motion intensity and EQI overtime. We observed that both motion intensity and EQI have larger variations during math and free time periods. However, the correlation between the EQI and motion intensity was only around 0.17 (Pearson Correlation, $p < 0.5$), which suggests that motion might not be the only source of noise artifact for the PPG sensor. The solid red vertical lines mark the timestamp when the signal processing estimator failed to produce RR measurements when a data window contains no valid sequence after the artifact elimination. In the bottom graph of Figure 2.10, we show the ground truth and the output from the signal processing and deep learning estimators. Since estimations from RIIV outperforms those from RIAV and RIFV (see Section 2.5), we only displayed the output with RIIV here. During the video period, we had fewer motions with lower EQI, and the signal processing outputs (RespWatch_RIIV) were closer to the ground truth than the deep learning outputs (RespWatch_DL). This shows that the signal processing estimator achieved high accuracy in the presence of moderate noise artifact (as indicated by the low motion intensity and EQI). However, when the motion or EQI increased, e.g., during math or free time periods, the signal processing estimator produced larger errors than the deep learning estimator. This demonstrates that deep learning estimator is more robust against higher level of noise, and highlights the advantages of our hybrid approach that utilizes the signal processing for high accuracy when EQI is low, and deep learning for robustness when EQI is high.

2.5 Evaluation of RespWatch

This section presents an evaluation of the three estimators supported by RespWatch. The accuracy of RR measurements was assessed using the mean absolute error (MAE) in breaths

per minute (bpm), defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_{ref,i}| \tag{2.5}$$

where n is the number of data windows, $\hat{y}_i$ is the estimated RR and $y_{ref,i}$ is the reference RR from ground truth. Moreover, we analyze the trade-off between accuracy and yield of RR measurements.

2.5.1 Signal Processing Estimator

We compared the performance of our signal processing estimator to existing methods for measuring RR based on PPG.

We implemented three state-of-the-art methods from [85, 150] as the baselines for performance evaluation. Those methods were previously evaluated on large data sets, and have been adapted to work with wrist-worn PPG [110]. The first two baseline methods were the *simple fusion* and *smart fusion* methods from [85], which utilized Fourier transform and fusion techniques. The third baseline method [150] utilized autoregressive (AR) models. All the three baseline methods have the data admission controls, which discard an entire data window that are found to contain noise artifacts. Table 2.1 shows the performance of the baseline methods on our dataset. We observed that a large portion of data windows were discarded due to the admission control. Unlike the baseline methods, our signal processing estimator employs fine-grained artifact elimination and can estimate RR even with some artifacts in a data window. We only discarded 13.86% of the data windows that contained no valid sequence after the artifact elimination. Hence, our signal processing estimator achieved a significant higher yield of 86.14% than the baseline methods.

Table 2.1: Baseline methods on our dataset

Method	Yield	MAE (bpm)
Karlen (2013) (Simple Fusion)	14.95%	1.876
Karlen (2013) (Smart Fusion)	11.87%	1.603
Pimentel(2017) (AR models)	14.29%	1.704

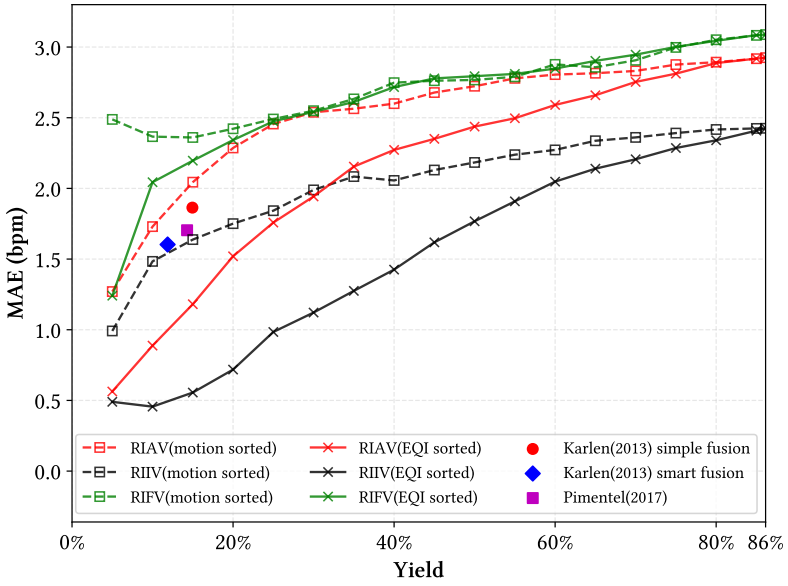


Figure 2.11: MAE vs. Yield. Different colors represent the estimations from RIAV, RIIV and RIFV, respectively. The line styles indicate different sorting criterion (Motion, EQI). The baselines are illustrated as dots with different shapes and colors.

We analyzed the trade-off between accuracy and yield of RR measurements. Figure 2.11 shows the *MAE-yield* curves for the signal processing estimator. As motions are previously used as a metric to reject the PPG measurements [140] and the proposed EQI is also capable for the same propose, we rank the data windows based on the corresponding motion intensity and EQI, respectively, to investigate the accuracy at different yields. For motion intensity, we computed the motion intensity for each PPG data window, and sorted the data windows by ascending motion intensity. Then, we calculated the MAE of RR measurements with motion intensity in the lowest α-*th* percentile, ranging from 5% to 100%. Similarly, we also calculated

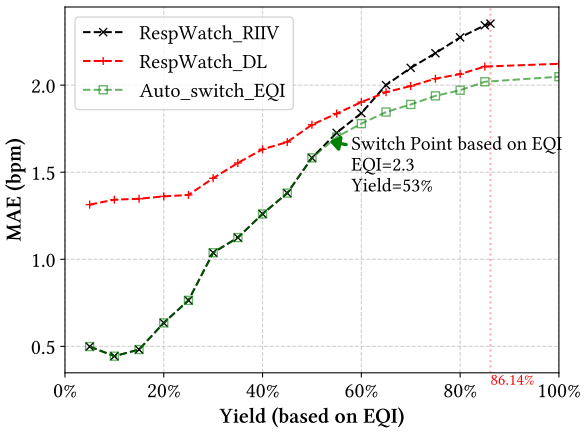


Figure 2.12: MAE vs. Yield based on EQI ranking.

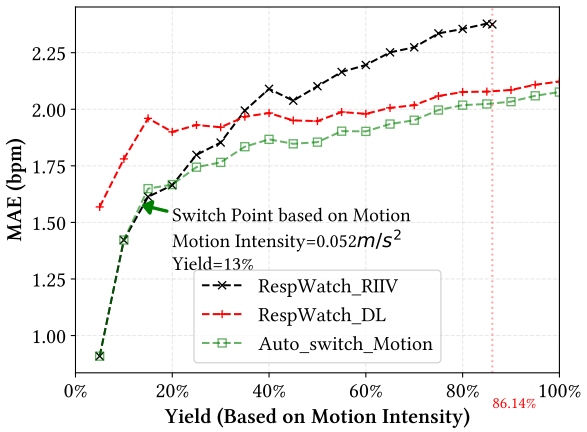


Figure 2.13: MAE vs. Yield based on Motion ranking.

the MAE for different yields using EQI as the metric for sorting the data windows. In Figure 2.11, Different colors distinguish the estimations from different respiratory-induced variation signals (RIAV, RIIV, RIFV), and different line styles distinguish the different ranking criterion (Motion intensity or EQI). For example, the dashed lines with square markers are the MAE curves with increasing motion intensity. Each point (α, e) on the dashed lines indicates the MAE of e on the subset of data windows whose motion intensities are in the lowest α-th percentile, corresponding to the yield of $\alpha\%$. The max yield of the signal processing estimator is 86.14% because we discarded 13.86% of the data windows that contain no valid sequence. In contrast, as the baseline methods have fixed data yield due to their data admission control policies, the results of the baseline methods are displayed as three discrete data points in the figure.

Observing the dashed lines with square marks in Figure 2.11, we found that all the three outputs from the signal processing estimator, RIAV(motion sorted), RIFV(motion sorted) and RIIV(motion sorted), had an increasing trend, suggesting that the motions indeed have a negative influence on the estimation accuracy. In practice, we may select the motion intensity threshold to achieve the desired trade-off between accuracy and yield of RR measurements

based on Figure 2.11. RIIV(motion sorted) outperforms all the baselines when at the same yield level. And RIIV(motion sorted) has the lowest MAE at any yield level among the three outputs from the signal processing estimator. This suggests that the RIIV is the most suitable respiratory-induced variation signal to estimate RR from the smartwatch PPG signals.

Next, we investigate the relations between the EQI and the accuracy of RR measurements. We note that RIIV(EQI sorted) significantly outperformed the three baseline methods, achieving around 3-fold decrease in MAE for the same yield, and also around 3-fold increase in yield for the same MAE. The solid curves in Figure 2.11 shows that the MAE also increases with EQI. However, the three solid lines are below the corresponding dashed lines with same color, especially for the RIIV and the RIAV. This indicates that when targeting the same yield, using EQI to reject PPG data can have lower MAE than using motion intensity. In another viewpoint, when targeting at the same accuracy, using EQI as the criterion to reject data can have a higher yield. Therefore, EQI is a more accurate indicator of measurement quality than motion intensity, as noise artifacts may be caused by sources other than motion.

The above evaluations demonstrated that our signal processing estimators can provide the flexibility to balance accuracy and yield according to the application requirements. Since the RIIV shows the best result, we focused on the signal processing estimator with RIIV for the following evaluations.

2.5.2 Deep Learning Estimator

In this subsection, we compared the deep learning estimator and the signal processing estimator. The deep learning estimator directly learns from the waveform and does not rely on any admission control or artifact elimination, so it produced estimations for all data windows, achieving 100% yield. We employed a 5-fold sample-based cross validation (CV)

scheme to train and test our deep learning estimator. We ensured there is no PPG waveform overlap between the training and testing set. The out-of-sample error is reported in the evaluation. Figure 2.12 and 2.13 plot the *MAE-yield* curves of the different estimators when the RR measurements are sorted based on EQI and motion intensity, respectively. The EQI is from the signal processing estimator with RIIV. For those 13.86% of data windows that signal processing estimator cannot estimate RR, we assigned an EQI of infinity. We observe that the signal processing estimator (RIIV) achieved lower MAE than the deep learning estimator when the EQI or motion intensity are lower. However, as EQI or motion intensity increases, the deep learning estimator becomes more accurate, suggesting a higher level of robustness against noise artifacts. The MAE dynamic range of deep learning is also not as large as it of signal processing, indicating deep learning is less sensitive with varying noise artifacts. The crossing point of the signal processing and deep learning curves in Figure 2.12 is at yield of 63%, while it is only at yield of 37% in Figure 2.13. And the deep learning curve in Figure 2.12 is relatively smooth compared to it in Figure 2.13. These once again show that the EQI can indicate the accuracy for RR measurements more accurately and smoothly than the motion intensity even for deep learning estimations.

2.5.3 Hybrid Estimator

For the evaluation of our hybrid estimator, we report the outputs dynamically chosen from signal processing and deep learning based on the best switching point. The best switching point of either EQI or motion intensity was obtained offline via the grid search. For real use, the hybrid estimator automatically switches between the signal processing and deep learning according to the best switching point without human efforts.

We first evaluated the EQI as the switching criterion. The green curve in Figure 2.12 shows the results of the hybrid estimator with EQI ($Auto\_switch\_EQI$). It achieves the best MAE

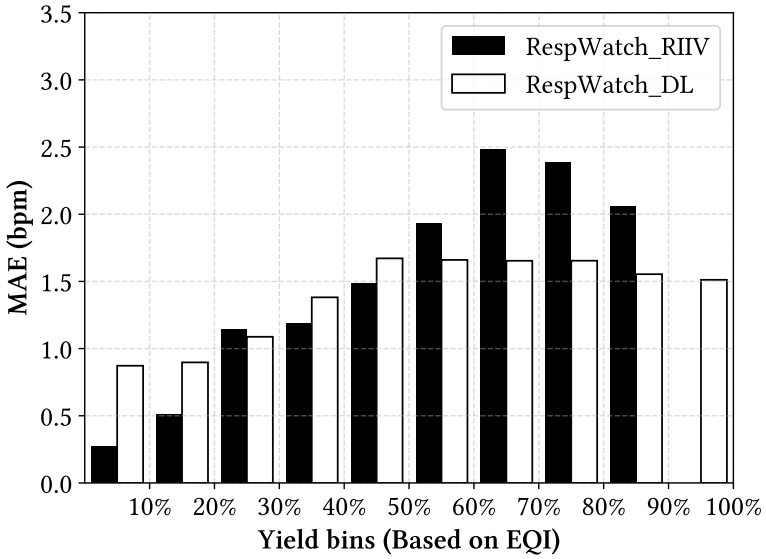


Figure 2.14: MAE in different yield bins with the EQI sorting criterion.

(2.017 bpm) compared to both deep learning and signal processing. The best switching point based on the grid search is around $EQI = 2.3$, corresponding yield of 53%, which means that the hybrid estimator automatically chooses signal processing when $EQI \leq 2.3$, and chooses deep learning when $EQI > 2.3$. We further investigated the relationship between MAE and EQI for the signal processing and deep learning estimators, as shown in Figure 2.14. The data windows were still sorted with the increasing EQI. Each bin contains the data windows with EQI from α-th to $(\alpha + 10)$-th percentiles. We observe that the MAE of $RespWatch\_RIIV$ becomes significant higher than $RespWatch\_DL$ from the sixth bin, corresponding EQI of range $[2.01, 2.45]$. The grid search was in a finer granularity, so we found the best switch point of EQI at round 2.3.

Besides, we explored the hybrid estimator switching with motion intensity, using the same grid search approach to find the best switching point of motion intensity. The green curve in Figure 2.13 shows the results. The hybrid estimator switching with motion intensity demonstrates slightly higher MAE (2.076 bpm vs. 2.017 bpm) than switching with EQI. And

it jumped to deep learning earlier at the yield of 13%, which utilizes significantly more times of deep learning outputs. This makes the hybrid estimator with the motion intensity less efficient. As a result, EQI is a better switching criterion in terms of both the accuracy and efficiency. So, in real implementation of RespWatch, we developed our hybrid estimator with the $EQI = 2.3$ as the switching threshold.

2.6 System Experimentation

2.6.1 Implementation on Smartwatches

Table 2.2: Information of the testing smartwatches

Device	Platform	RAM	System	PPG Sensor
Fossil Gen4	Wear 2100[1]	512MB	H	PAH8011[2](200Hz)
Fossil Sport	Wear 3100[3]	512MB	H	PAH8011 (100Hz)

We have implemented RespWatch on Wear OS in mainstream smartwatches. Wear OS [50] is a version of Android operating system tailored for smartwatches and other wearables. For the CNN model in the deep learning estimator, we first trained the model on the server in PyTorch [142] framework, and then transcripted the model into mobile version[153] on Wear OS. For the hybrid estimator, we chose the switching scheme of $EQI = 2.3$ based on our results in Section 2.5.3.

[1]https://www.qualcomm.com/products/snapdragon-processors-wear-2100
[2]https://www.pixart.com/products-detail/27/PAH8011ES-IN
[3]https://www.qualcomm.com/products/snapdragon-wear-3100-platform

2.6.2 Empirical Evaluation

Two smartwatches were used in our empirical evaluation, shown in Table 2.2. Each experiment was repeated 500 times, and the average of running time and resource usage are reported in Table 2.3 and 2.4.

Table 2.3: Profile of Signal Processing Estimator

Devices	Preprocessing	Art. Elim.* & Pulse Peak Finding	RIXV* Extraction & Adaptive Peak Finding	Total Time	Ave. CPU(%)	Energy
Fossil Gen4 (H)	5.836ms	19.139ms	19.919ms	44.895ms	53.53%	Light to Medium
Fossil sport (H)	5.385ms	16.058ms	16.621ms	38.064ms	50.25%	Light to Medium

*Art. Elim.: Artifact Elimination
*RIXV: Respiratory-Induced Variations (RIAV, RIIV, RIFV).

Table 2.4: Profile of Deep Learning and Hybrid Estimator

Devices	Prep. (ms)	CNN (ms)	Deep learning Total Time (ms)	Ave. CPU (%)	Energy	Hybrid with EQI*	Hybrid with Motion Intensity*
Fossil Gen4 (H)	8.856	6504.262	6592.828	85.34%	>Medium	2879.811ms	5780.655ms
Fossil sport (H)	8.472	7934.962	7943.434	70.23%	∼=Medium	3453.740ms	6948.851ms

*The running time of hybrid estimator is the expected running time based on our dataset with the corresponding best switching threshold.

The signal processing estimator was highly efficient with a total running time less than 50 ms (see Table 2.3), whereas the deep learning estimator had a total running time higher than 6000 ms (as shown in Table 2.4). The average energy consumption and average CPU utilization were acquired through the Android Profiler [4]. The signal processing estimator consumes less energy with lower CPU utilization. For the deep learning estimator, the CNN model consumed about 98% of the total time, suggesting the need for optimization in the future. For the hybrid estimator based on the best switching threshold on our dataset, the expected running time of switching with EQI was significantly lower than switching with Motion Intensity.

Our results established the feasibility to run RespWatch locally on smartwatches for RR monitoring. Even though the deep learning estimator takes about more than 6 seconds to run per RR measurement, it only needs to be executed every 1 minute for a RR sampling rate of once per minute. The results also highlight our hybrid approach, which only invokes the deep learning estimator in the presence of significant noise artifacts with high EQI. The hybrid method drastically lowers the running time and saves energy while maintaining high accuracy.

2.7 Discussion

We have demonstrated our RespWatch in this study, which outperforms the state-of-the-art baseline methods. The empirical evaluations also quantitatively reveal the execution efficiency and capability of running on the commercial smartwatches. Nevertheless, there are still some room for future improvement.

RR measurement with excessive motions. Our RespWatch system has been tested in various activity scenarios involving motions, but it did not cover all scenarios in our daily life. The evaluations in this paper shows the feasibility and high accuracy of RespWatch in situations, like working in front of computer, resting and others activities with motions to some degree. The applicability of RespWatch, especially in scenarios with excessive motions (e.g. running) still needs to be evaluated.

Inter-individual Difference. For both the signal processing estimator and deep learning estimator, we applied the same parameters for all the subjects. The differences between individuals (e.g., skin tone and wearing habits) may have impacts on PPG signals. Earlier studies [74] show the personalized models may mitigate the impact. A potential research

direction is to leverage personalized models while minimizing the burden of the personalizing process.

Detection of Potential Respiratory Diseases. Currently, our evaluation is limited to healthy volunteers. Further studies are needed to test the feasibility of wearable RR monitoring for patients and exploit the technology to detect respiratory conditions.

2.8 Conclusion

RespWatch is a wearable sensing system for robust RR monitoring on smartwatches with PPG. We explored signal processing, deep learning and hybrid approaches to measure RR based on PPG signals. We collected a large dataset from 30 participants who performed multiple activities with the smartwatches. The signal processing estimator achieved higher accuracy in the presence of moderate noise artifacts, while the deep learning estimator was more robust to significant noise artifacts. Given the complementary strengths, we developed a novel hybrid estimator that can automatically switch between the signal processing and deep learning based on the EQI. The hybrid estimator not only achieved the best accuracy but also leveraged the high efficiency.

Chapter 3

Detecting Objective and Subjective Stress Using Smartwatches

This chapter introduces stress detection models using smartwatches. Unlike the respiratory rate measurements, stress is a more complex physiological or psychological response.

3.1 Introduction and Related Work

Stress is a common health concern and chronic stress is associated with the development of depression and anxiety [204], immune function dysregulation [122, 176], cardiovascular disease [112], decreased work-related performance [6], quality of life [169], and drug use [185]. In the United States, over 50% of the adult working population have described their work productivity being affected by stress [7]—resulting in diseases, absenteeism, presenteeism, and staff turnover. Such loss in productivity cost nearly 187 billion dollars [56].

The accumulated daily stress contributes to chronic stress. In spite of its significant impact, routine measurement of stress has been challenging. The complexities of measurement arise from difficulties in discerning appropriate physiological and subjective measures of stress, and the considerable intra- and inter-individual differences in the manifestations of stress [49]. Objective measurements have relied on the measurement of cortisol and inflammatory cytokines [69, 141] that have been used as successful objective proxies for measuring stress [96]. However, such measurements are not pragmatic in real-world, routine stress situations. The most widely used method for measuring stress is through questionnaires. Survey scales such as the Perceived Stress Scale (PSS) [28], and Depression Anxiety Stress Scales (DASS) [132] have been shown to be effective in measuring perceived stress in different cohorts [166, 186]. Although useful, these self-reported questionnaires are time-consuming, suffer from recall bias and provide only a snapshot view of an individual's perceived stress [49].

Newer approaches using mobile or internet-enabled devices with ecological momentary assessments (EMA) and participant self-reports, can considerably increase the sampling frequency, providing insights into individuals' activities, affect and behaviors [182]. Within this context, EMAs or self-reports afford a viable mechanism to detect and characterize the evolution and progression of stress [209]. However, even EMAs or self-reports with a high frequency of contact have shown to decrease the quality and response rate among participants over time [76, 209]. This is because EMAs and self-reports have been shown to be affected by recall bias, fabrication, and falsification in reporting [70].

The exponential growth and adoption of wearable technology have afforded new opportunities to measure and monitor a number of physiological signals including skin conductance, skin temperature, electrocardiogram (ECG) and photoplethysmogram (PPG). Physiological measurements derived from these sensors have also been used to develop machine learning-based prediction models [70, 94, 151, 175]. For developing these models, most of these studies

have relied on laboratory-based trials where artificial stress stimuli were induced. For example, Hovsepian et al. used a combination of ECG and respiration inductive plethysmography to predict stress using machine learning algorithms [70]. Similarly, King et al. focused on a cohort of pregnant women to develop stress models from laboratory-based studies and translated these models for "in the wild" studies [94]. Several other studies have also developed similar models using a combination of sensors relying primarily on laboratory-based studies [151, 188].

Although these models had relatively high performance in laboratory-based settings, there are several challenges. First, most of these studies used a combination of multiple body-worn sensors that are pragmatically difficult to translate for real-world clinical applications. The chest belts that were used in several of these studies [70, 94, 151], are cumbersome to use in free-living situations, limiting user compliance and data yield. Second, many of the machine learning models that were developed using the laboratory-based trials may not be applicable in free-living settings, where we would need to rely on participant self-reports rather than specific induced stress stimuli. To address these gaps, we had the following methodological and research objectives: first, to instrument a commercially available smartwatch for detecting and predicting stress in controlled scenarios. Such measurements with a commercial smartwatch—using both physiological measurements and associated self-reports—have translational potential for use in clinical settings. Second, to compare the objective markers of stress with participant-reported subjective markers of stress from self-reports. Such a comparison can help in establishing the viability of using self-reports as a potential proxy measurement mechanism for stress. Finally, to develop an approach for creating personalized stress models by accounting for inter-individual differences.

3.2 User Study

The user study was the same in Chapter 2. Here we elaborate on some details related to stress.

3.2.1 Participants

32 healthy volunteers were recruited through flyers posted across the campus at Washington University in St. Louis. Respondents who met the inclusion criteria—between 18 and 69 years of age, with no heart disease, not pregnant at the time of recruitment, and not having an implanted pacemaker—were screened over the phone; if participants met all inclusion criteria, they were recruited for the study. All participants received a $25 Amazon gift card. The institutional review board of Washington University approved this study, and written consents were obtained from all participants (IRB#2019-04150).

3.2.2 Study Design and Procedure

Participants completed an approximately 2-hour laboratory-based phase and a 24-hour field-based phase. This study mirrors previous experimental designs utilized for detecting stress patterns using wearable sensors [70, 175].

In the laboratory-based phase, recruited participants first completed two surveys on paper: the 10-item Perceived Stress Scale (PSS) [28] and the 42-item Depression Anxiety Stress Scales (DASS) [132]; both of these surveys have been shown to be effective in the measurement of perceived stress and depression [27, 136]. After the completion of the surveys, participants were asked to wear the Fossil Gen4 Explorist (see Figure 3.1 (B)) smartwatch on their non-dominant hand.

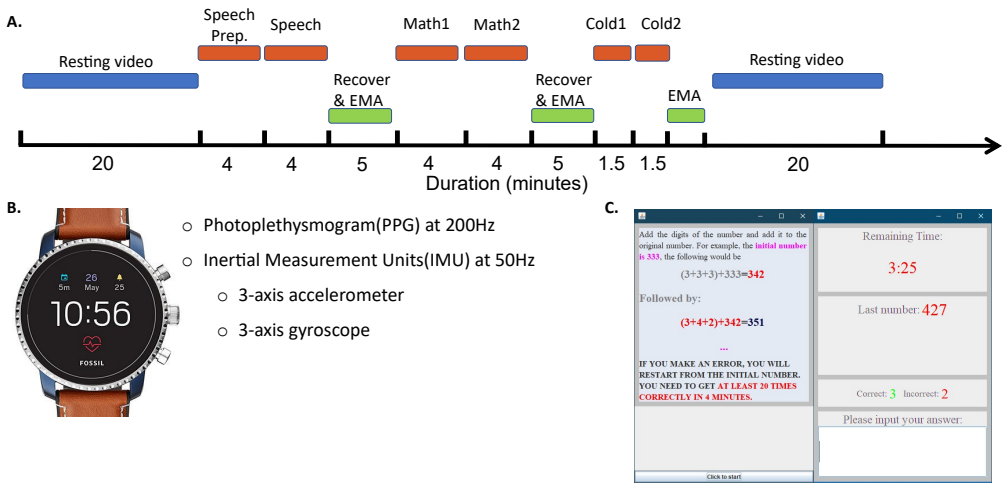


Figure 3.1: Study procedures and devices

The laboratory-based phase included several stages (see Figure 3.1 (A) for the sequence). First, participants had a 20-minute resting period, during which they watched a relaxing nature-oriented program, as a "non-stressed" period [70, 136, 151] (we refer to this period as "video-based resting period"). During this period, participants were left alone in a room and asked to relax as much as possible. If participants felt stressed or uncomfortable during this video-based resting period, they were instructed to discuss with the study coordinator to potentially stop their continued participation. No participants withdrew during this video-based resting period.

After the first 20-minute resting period, participants were provided with general instructions regarding a series of tasks related to public speaking, mental arithmetic, and cold stressor. These tasks corresponded to social, cognitive, and physical stressors and have been applied as stress-inducing stimuli in controlled settings in previous laboratory-based studies [70, 87, 151]. Participants completed each of these stressor tasks in the same order, with 5-minute resting and recovery periods between each stressor task. We asked the participants to hold their watch-wearing hand as still as possible during the laboratory-based phase, as the physiological signals recorded from the smartwatch are vulnerable to physical motion [187].

The first task was the public speaking task. For this task, participants were given a topic, a 4-minute preparation time, and then were asked to speak to the study coordinator and a researcher in the room for a period of 4 minutes. At the end of the public speaking task, participants were given a 5-minute rest and recovery period. Next, participants were given instructions regarding a mental arithmetic task. This task was completed on a computer using an application that we developed (see Figure 3.1 (C)). The task involved mentally adding the digits of a number, and then adding the total to the original number. For example, if the initial number was 234, the sum of the digits was 9, and the next number would be 243. On the application, there was a countdown timer (for 4 minutes), and an indicator for the number of errors that the participant made. If a participant made three consecutive errors, they had to restart the task. Participants were asked to achieve at least 20 accurate responses. Participants completed the arithmetic tasks sequentially twice, once standing up, and once seated on a chair, in the same order. At the end of the arithmetic task, participants were given another 5-minute rest and recovery period.

The last task was the cold stressor task. We used a custom-made solid stainless-steel cylinder (10 centimeters high, 5 centimeters diameter) that is routinely used for cold allodynia tasks [206]. The rod was kept in the refrigerator for a period of 12 hours and measured approximately at 4℃ at study time. The task involved each participant holding the rod for a period of 90 seconds on each hand (first, in their dominant hand, followed by the non-dominant, in that order). If the pain was unbearable, participants were asked to release their hold prior to the end of the testing period (i.e., 90 seconds). At the end of the cold stressor task, participants were asked to watch another nature-oriented program for a period of 20 minutes as a relaxation period.

As previously described, at the end of each stressor task, there was a 5-minute resting and recovery period that potentially allowed for the stressor to subside prior to the next exposure.

Additionally, at the end of each stressor task, participants were automatically sent a self-report on the watch that asked them regarding their "stress" with a 4-option response: "[Happy] [Stressed] [Tired] [Neutral]." This self-report was based on scales used in previous studies on the measurement of stress [70, 94, 151]. For example, Zachary et al. showed that the "Happy" response was negatively correlated to the intended stress, which can be used as an indicator of non-stress [94]. The "stressed" and "neutral" responses were direct indicators of self-reported (i.e., subjective) stress and non-stress.

At the completion of the final resting and recovery period, participants were given instructions for the field-based phase on wearing the watch, charging, and responding to self-reports. In addition, participants were given the smartwatch, its charger, and a paper-based physical activity tracker. Participants recorded their physical activity that they participated in while wearing the watch along with the start and finish times on the paper-based physical activity tracker. Additionally, all participants had a return date and time scheduled at the end of the field phase such that all materials could be collected. At the return visit, all participants were given a $25 Amazon gift card for their study participation. The data collected in the field phase was exploratory to evaluate the viability of collecting self-reports and collecting physiological data regarding stress in free-living situations.

3.3 Method

3.3.1 Definition of Subjective and Objective Stress

As previously described, we used two descriptors for categorizing the signals drawn from the smartwatch. First, we used the signals drawn from the stressor tasks (social, cognitive, physical) as "stress periods", categorizing them as objective stress. We defined objective stress

as the biological reaction to a stressful exposure that manifests with biological reactions or changes (e.g., changes in cortisol levels) [100, 177]. Second, we used the self-reported responses from participants (categorized as stressed, not stressed) as the labels for categorizing the stressor tasks. We call this subjective stress or a subjective feeling of "being stressed" [49].

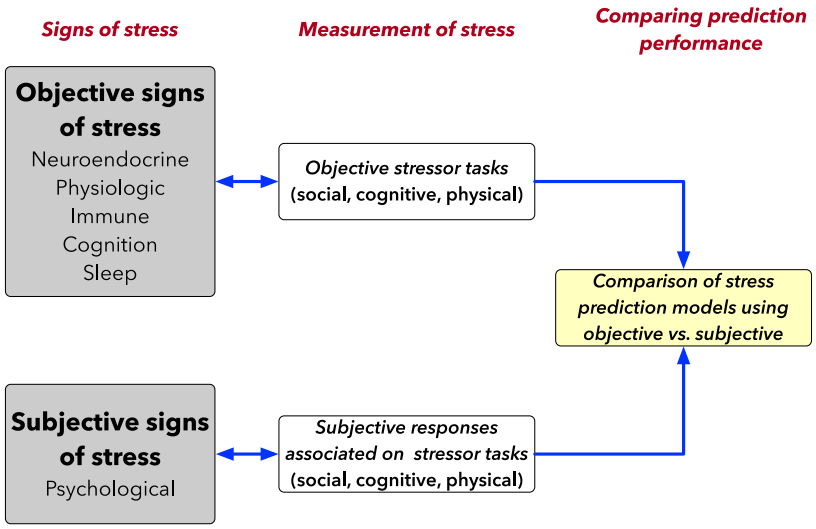


Figure 3.2: Comparing objective stress and subjective stress.

Based on a participant's self-reported response after a task, we labeled the preceding task as "stressed" or "not stressed." Our framework for analysis compared the performance of stress prediction models developed using objective stressors and the subjective responses (i.e., self-reports) for each of the stressor tasks (see Figure 3.2). This approach established the potential for using subjective responses as a proxy for stressors.

3.3.2 Data Preprocessing

We conducted several data pre-processing tasks to translate the raw data file into interpretable physiological readings. This preprocessing pipeline was based on the RespWatch (see Chapter 2). Across all study participants, there was an average sampling rate of 206.02 Hz for the

PPG sensor and 48.81 Hz for the IMU sensor. As the sampling rates were not stable, we synchronized data based on the timestamps of each sensor event and then re-sampled the synchronized PPG and IMU data at 200Hz with Hermite spline interpolation [129]. The re-sampled data were segmented into sliding windows with a window size of 60 seconds and a step size of 20 seconds. The 60-second window size has been previously used for stress-related studies [70, 94].

Intense movement and poor physical contact between the PPG sensor and skin can potentially degrade the quality of the PPG signal. Towards this end, we first employed a forward-backward Butterworth bandpass filter to remove the noise outside of heart rate and respiration band with a cutoff frequency of 0.15Hz and 4Hz in each window [44]. To screen out motion artifacts and poor signal sequences within the heart and respiration band, we further utilized a sliding sub-window approach, which divided each 60-second window into 10-second sub-windows with a 2-second step size. The motion detector [54] and heartbeat pattern detector [135] were then applied on each of these 10-second sub-windows. The motion detector detects movement based on the IMU sensor. The heartbeat pattern detector can validate whether a PPG waveform matches a valid heartbeat pattern.

Only the sub-windows that passed both detectors were marked as valid signals. Once we iterated through all sub-windows within the 60-second window, valid consecutive sub-windows were merged into a larger "valid" segment. Features were extracted only from these valid segments of signals. This approach helped in eliminating short invalid signal periods within 60-second windows and increased the availability of testing samples (see a similar approach in [94]). Figure 3.3 shows an example of how we translated the raw PPG signal into "valid" and "invalid" segments. We set a threshold of 25% for the invalid period. If the invalid period was more than this threshold, the entire 60-second window was discarded.

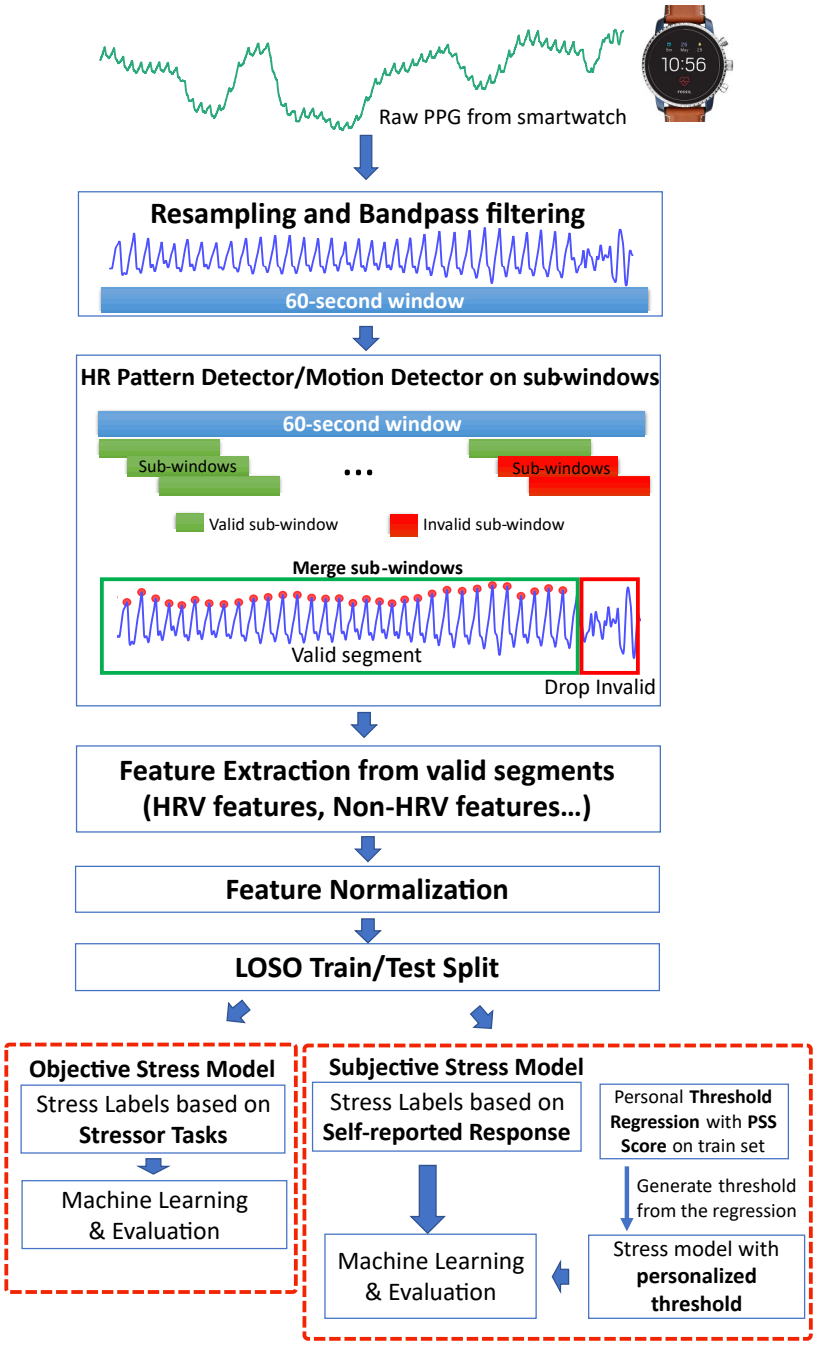


Figure 3.3: Overview of the multi-stage data processing, machine learning for objective and subjective stress machine learning pipeline. (LOSO: leave one subject out.)

We also retrieved self-reported responses from participants at the end of each stressor task.

The self-reported responses were grouped into two categories: "stressed" (with responses of

stressed) and "not stressed" (with responses of happy, tired, or neutral). These responses were used as labels for the analysis for determining subjective stress patterns (see Section 3.3.1).

3.3.3 Feature Extraction

Features were extracted from valid segments in the 60-second window after pre-processing. We used the peak detection method with adaptive filtering [34] to extract the inter-beat interval (IBI) series from PPG data. IBI refers to the time interval between individual heart beats. Figure 3 displays the waveforms from the PPG sensor. Each peak (marked by a red dot) represents one heartbeat. By detecting the peaks of the waveform and calculating the horizontal distances between the dots, we can extract the time series of the IBI.

Heart Rate Variability (HRV) features and other non-HRV features were derived from the IBI series. We also used the Detrended Fluctuation Analysis (DFA), a method to measure the statistical self-similarity of a signal, to determine non-stationarity within the IBI series [146].

In addition, we extracted respiration-related features, which have been known to be associated with stress [66, 70]. Though the smartwatch sensors do not directly provide respiratory rate, we estimated respiratory rate based on three respiratory-induced variations from the IBI series and the PPG signal, similar to what has been used in previous research [80, 85]: respiratory-induced amplitude variation (RIAV), respiration-induced intensity variation (RIIV), and respiratory-induced frequency variation (RIFV). RIAV is the change in peripheral pulse strength, caused by reduced ventricular filling, and is the peak height change in the raw PPG waveform. RIIV is the change of perfusion baseline, caused by the intrathoracic pressure, and is the intensity change in each peak-to-valley in the raw PPG waveform. Finally, RIFV is the change of heart rate, caused by the autonomic response to the respiration cycle, and is represented as the heartbeat interval change. We interpolated these three variations using

linear interpolation at 100Hz. Since the respiration is cyclic, the respiration-induced variations could be also cyclic. We employed Fast Fourier Transfer (FFT) with a Hamming window to calculate the major frequency for each respiration rate variation in the data window. The major sinusoidal frequency in the FFT was used as an estimation of the respiratory rate.

Table 3.1: Features that were extracted from the IBI and PPG signals.

Category	Features
HRV features	SDNN, RMSSD, SDSD, pNN20, pNN50, low frequency (LF) energy (0.04-0.15 Hz), high frequency (HF) energy (0.15-0.40 Hz), LF/HF energy ratio (LF_HF)
Non-HRV IBI features	mean, median, minimum, maximum, interquartile range (iqr), 20th percentile, 80th percentile, detrended fluctuation analysis (DFA), heart rate
Respiration-related features	FFT_RIIV, FFT_RIAV, FFT_RIFV

HRV features [180]: SDNN, standard deviation of the IBI of normal heartbeats; RMSSD, root-mean-square of successive differences between normal heartbeats; SDSD, standard deviation of differences between adjacent IBI; pNNX, percentage of successive IBIs that differ by more than X milliseconds. Non-HRV features are other features extracted from the IBI time series. Respiratory-related features: FFT_RIIV, FFT_RIAV, FFT_RIFV extracted with the RespWatch-based pipeline and FFT.

A total of 20 features were extracted (see Table 3.1). All features were standardized using a normalization method, where the median was removed, and each feature was divided by its interquartile range. As there are large differences in an individual's physiological signal manifestations, we applied this normalization method on each individual's feature data to alleviate the subject-specific components in the feature data [70, 94, 189].

3.3.4 Model Training and Validation

We focused on the machine learning analysis during the lab phase, as the primary goal was to compare the objective and subjective stress detection performance. The presence of well-defined "ground truth" data—the objective stressor tasks and subjective self-reports—during this phase enabled such a focused analysis.

We used multiple machine learning models for both the objective and subjective stress detection including support vector machine (SVM), random forest (RF), AdaBoost, gradient boosting decision trees(GBDT) and logistic regression (LR). These models have been widely applied in the literature on developing similar health-related models [12, 94]. These models also generate probability estimates for each prediction. By tuning the threshold probabilities, it is possible to achieve a desired sensitivity or specificity. We used a fixed threshold for the prediction, where a signal with a probability$>$0.5 was categorized as stressed.

The hyperparameters for each model were tuned with grid search to achieve the highest F1-Score. F1-score is the harmonic mean of the precision and recall [8]. When training the model, we up-sampled the minority class to avoid skewed prediction on the majority class, as we have more data in the resting period (non-stressor) than in the stressor tasks. We applied leave-one-subject-out (LOSO) cross validation; in other words, we evaluated each participant's data with a model trained on all the other participants' data. This ensured that there was no overlap for each participant between the training and validation dataset. After choosing the model with the best F1-score, we ran the feature selection algorithms to eliminate the highly correlated [52] and unimportant features. For the SVM model with radial basis function (RBF) kernels, we employed the multi-kernel learning for feature selection. The feature importance was ranked based on the kernel weight coefficient for each feature [161]. For the tree-based models, the feature importance was derived from the Sklearn Python package

[1]. For the LR, the weight coefficient of each feature was regarded as feature importance. Feature selection helped in trimming the model and avoiding overfitting.

For both objective and subjective stress models, we computed the area under receiver operating curve (AUROC), accuracy, sensitivity (recall), specificity, and precision (positive predictive value). All the evaluations were run 10 times, and an average performance metric with standard deviations was used for all reported results. In the machine learning models described above, we first used the same threshold of probability estimates (=0.5) for each participant. In other words, we classified data signals as stressed if the probability estimates exceeded this threshold.

3.3.5 Personalized Subjective Models with Adaptive Threshold

The subjective self-reported response usually suffers from individual differences in responses [193], as it is "subjective" to individuals. To address the challenge induced by those inter- and intra-individual differences in the experience of stress, previous literature [70, 151] had investigated training separate machine learning models for each individual. Nonetheless, this approach requires a complex individual training process before the deployment, and could potentially incur overfitting issues. To handle the inter-individual differences and alleviate the burden of training personalized models for each person, we proposed a novel personalized subjective model with an adaptive threshold. We observed that the mean PSS score of people who reported being stressed was larger than those who reported not being stressed, for both the social and cognitive stressors. Although the differences between the non-stressed and stressed groups were not statistically significant (see Table 3.2), they could be possibly used to guide subjective stress prediction models. As such, we incorporated a personalized threshold into the general subjective stress model, by exploiting the correlation between the model prediction threshold and the participant's pre-study PSS score.

Table 3.2: Differences in the pre-study PSS scores for participants reporting as stressed or non-stressed with each of the stressor tasks (on self-reports).

	PSS score	
	Difference of mean	p value
Social (Speech)	3.14	0.085
Cognitive (Math)	3.18	0.055
Physical (Cold)	1.67	0.243

First, we trained a general subjective stress model based on the self-reported response. Then, we used a grid search to extract the threshold achieving the best prediction accuracy score of subjective stress detection for each participant in the training set. Based on the best threshold, we used ridge regression [58] to fit the relationship between the best threshold and a participant's pre-study PSS score. The threshold for each individual in the testing set was generated from the regression model, and we used this threshold to classify signals as stressed or as not stressed for the individual. This approach personalizes the stress detection with the generated threshold from the PSS score regression model, while avoiding the complexity of training a personalized machine learning model for each individual (See Figure 3.4 and Figure 3.3).

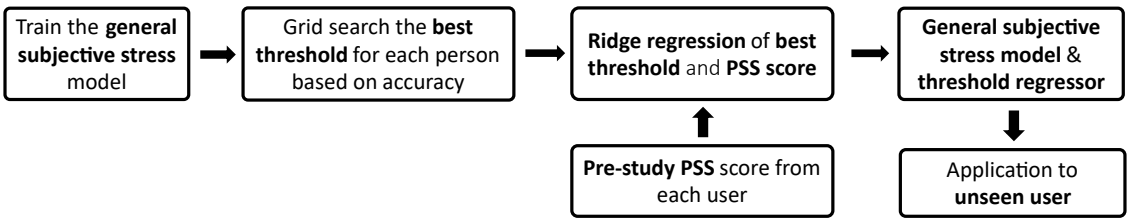


Figure 3.4: Workflow of the personalized subjective stress detection model

3.4 Evaluations

3.4.1 General Characteristics

All participants (n=32) successfully completed the entire study protocol. Participants were primarily female (n=24), with an average age of 36 years (S.D.=12.6). Except for nine participants who only completed part of the cold stressor tasks, all other tasks were completed by all participants. In total, we captured 1700 minutes of PPG signal data, with a 1160-minute resting period, and a 540-minute stressor period. However, data on two participants were not used for final analysis due to a partial malfunction of the smartwatch.

The average PSS score across participants was 11, showing the participants had average low or moderate perceived stress (S.D.=5.0); based on the PSS score range [27, 128, 147], 10 participants reported moderate stress, 22 participants reported low or no stress, and no participants reported high stress. The average stress score on the DASS scale was 5.2, (S.D.=4.5). A moderate positive relationship was observed between the PSS and DASS ($r=0.53$, $p<0.005$).

3.4.2 Predicting Stressed and Non-Stressed Periods from Stressor Tasks (Objective Stress)

We first investigated the ability of machine learning models to predict objective stress. Based on the objective stress definition, we labeled the data during the stressor tasks as "stressed" and the video-based resting period as "not-stressed". A fixed threshold of 0.5 was adopted for the probability output from the machine learning models: i.e., if the probability was greater than 0.5, we classified the signal as stressed (and vice-versa).

We found that the SVM outperforms other machine learning models, with an F-1 Score of 0.623, highlighting that these models have predictive capabilities of differentiating stress and resting periods (see Table 3.3)

Table 3.3: Predictions of stressed and non-stressed periods using multiple machine learning algorithms for objective stress. Mean (S.D.) are reported.

Model	Precision	Recall	Specificity	F-1 score	Accuracy	AUROC
SVM[1]	.625(.023)	.621(.010)	.888(.011)	.623(.014)	.826(.073)	.790(.007)
RF[2]	.598(.030)	.632(.011)	.872(.012)	.614(.013)	.817(.074)	.804(.013)
AdaBoost	.472(.012)	.641(.012)	.785(.013)	.543(.006)	.750(.072)	.749(.015)
GBDT[3]	.528(.011)	.652(.010)	.825(.006)	.584(.009)	.784(.072)	.790(.011)
LR[4]	.516(.008)	.649(.012)	.817(.008)	.575(.005)	.776(.077)	.772(.014)

[1]Support vector machine; [2]Random Forest; [3]Gradient Boosting Decision Trees; [4]Logistic regression.

Next, we investigated whether the machine learning models can differentiate between the three induced stressor tasks, i.e., the social (speech), cognitive (math), and physical (cold) stressors. As the SVM achieved the best performance for differentiating stress and non-stress periods, we evaluated its performance on each of the stressor tasks.

We found that the performance of the SVM model was highest for the social (F-1 score=0.685), followed by the cognitive stressor (F-1 score=0.610). The cold stressor had the worst performance (F-1 score=0.202), suggesting the difficulty to differentiate the cold stressor task from a resting state, using features derived from PPG data (See Table 3.4).

We investigated the potential causes for the lower performance of the physical (cold) stressor tasks. Using t-distributed stochastic neighbor embedding (t-SNE), an unsupervised approach, we visualized the extracted features across the three stressor tasks (See Figure 3.5). The t-SNE plot showed that features from the social and cognitive stressor tasks (yellow and red dots) were separated from the large cluster of resting tasks (dark green dots). We can observe

Table 3.4: Predictions of social, cognitive and physical stressor tasks using the SVM model. Mean (S.D.) are reported.

Stressor Task	Precision	Recall	Specificity	F-1 score	Accuracy	AUROC
Social (Speech)	.634(.023)	.747(.033)	.967(.004)	.685(.020)	.951(.004)	.963(.006)
Cognitive (Math)	.541(.025)	.700(.033)	.956(.004)	.610(.023)	.940(.004)	.933(.006)
Physical (Cold)	.124(.037)	.740(.053)	.745(.004)	.202(.046)	.732(.004)	.744(.006)

a relatively clear separation boundary between the social and cognitive stressors and the video-based resting period, but no clear separation boundary between the physical stressor and the video-based resting period. This lack of a clear boundary or separation between the physical stressor and the resting period potentially explains the lower performance on the physical stressor.

3.4.3 Predicting Stressed and Non-Stressed Periods from Self-reported Responses (Subjective Stress)

For predicting the subjective stress, we used the participants' self-reported responses after each stressor task as the ground truth. When a participant's self-reported response was "not stressed," we labeled the data during that stressor task as "non-stressed" (and vice-versa). Nearly 48% stressor tasks were labeled as "stressed" based on participants' self-reported responses. We first used the fixed threshold of 0.5 as the cutoff to classify the stressed and non-stressed (same as the objective stress model).

Similar to objective stress, SVM achieved the highest F-1 score (0.520). However, the model performance was lower, indicating that the subjective stress was potentially harder for machine learning algorithms to detect.

We performed one-way ANOVA tests on all features (from Table 3.1) for each stressor task comparing self-reported responses of "stressed" and "non-stressed" (See Figure 5). Comparing

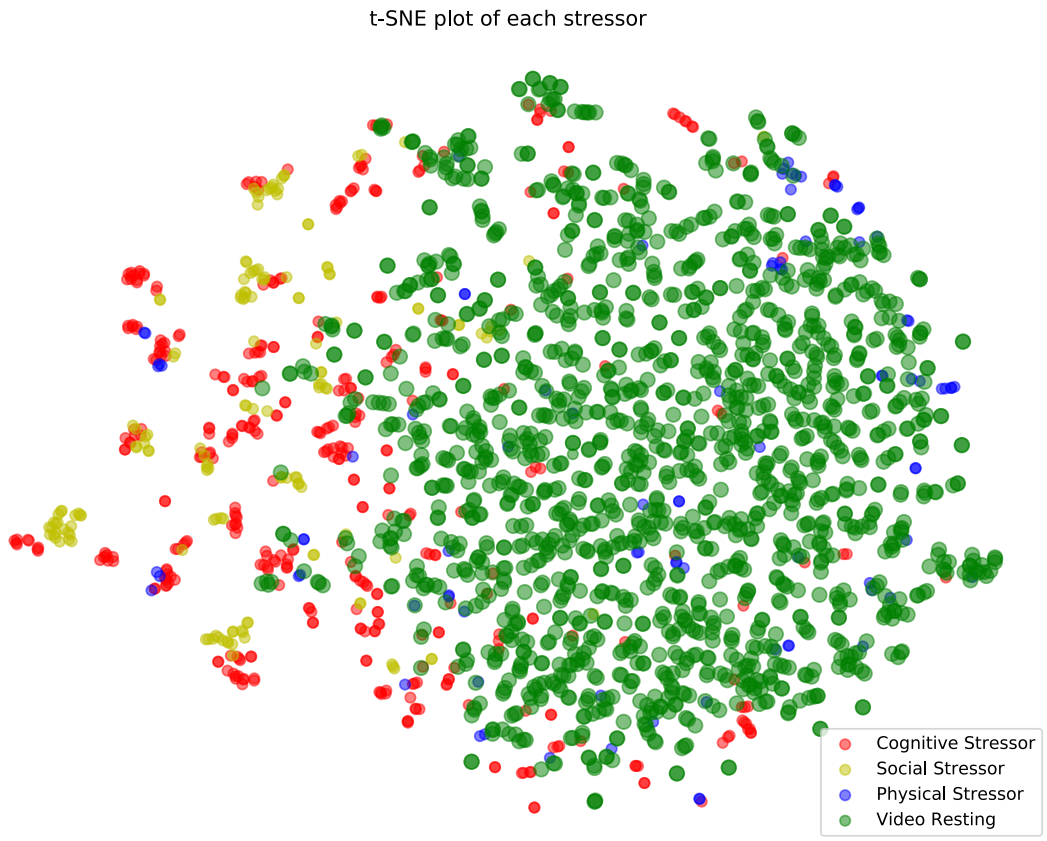


Figure 3.5: Clusters of various stressor activities based on t-distributed stochastic neighbor embedding (t-SNE).

Table 3.5: Predictions of stressed and non-stressed periods using multiple machine learning algorithms for subjective stress. Mean (S.D.) is reported (based on the fixed threshold of 0.5).

Model	Precision	Recall	Specificity	F-1 score	Accuracy	AUROC
SVM[1]	.469(.015)	.584(.014)	.801(.011)	.520(.012)	.744(.102)	.719(.007)
RF[2]	.584(.030)	.408(.017)	.913(.007)	.480(.022)	.798(.099)	.726(.007)
AdaBoost	.458(.010)	.521(.014)	.815(.007)	.487(.011)	.744(.093)	.694(.013)
GBDT[3]	.508(.009)	.478(.014)	.861(.004)	.492(.010)	.771(.094)	.697(.009)
LR[4]	.433(.011)	.598(.009)	.765(.008)	.502(.010)	.717(.102)	.720(.006)

[1]Support vector machine; [2]Random Forest; [3]Gradient Boosting Decision Trees; [4]Logistic regression.

each stressor with the video-based resting (i.e., social stressor-resting (S-R), cognitive stressor-resting (C-R), physical stressor-resting (P-R)), we found that the HR during social and

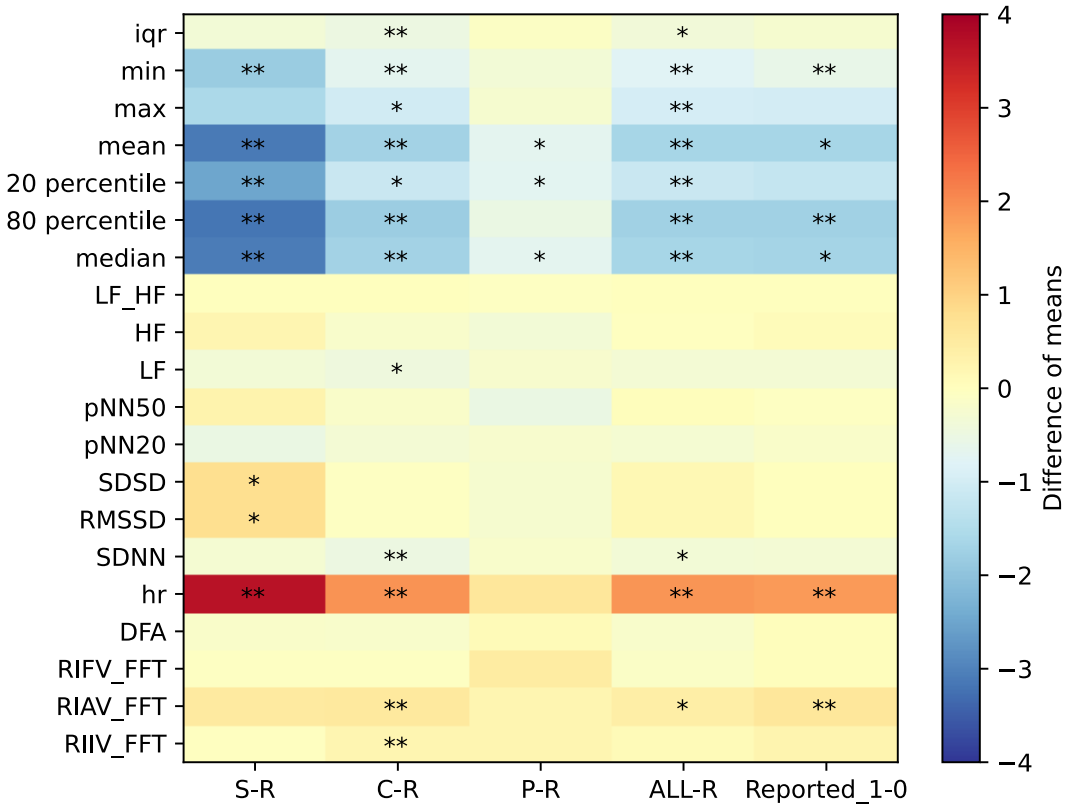
cognitive stressors was significantly higher than during video-based resting. The SDNN and LF were significantly lower during the cognitive stressor. SDSD and RMSSD were higher for the social stressor. Other HRV features did not show significant differences between the stressor and video-based resting.

Similarly, respiration-related features such as the RIAV_FFT and the RIIV_FFT were significantly higher during the cognitive stressor. Nonetheless, comparisons between the physical stressor and the resting phases showed fewer features with significant differences. Similarly, we found that periods labeled using self-reported stress (i.e., subjective stress) had fewer significant differences in the considered features compared with the objective stressor models (See the last column of Figure 3.6), which were consistent with their lower machine learning performance.

3.4.4 Predicting Subjective Stress with a Personalized Threshold

We used grid search to find the prediction threshold achieving the best accuracy for each participant. Table 3.6 shows the prediction results using the best threshold. The results have a significant improvement compared with the result with the fixed threshold (=0.5) (See Table 3.5). However, this method involved the ground truth to select the threshold, which is not generalizable to the unseen datasets.

To investigate the relationship between the PSS scores and the best probability threshold, we measured the degree of association using Pearson's correlation coefficient. There was a negative correlation between the PSS score and the best probability threshold with $r = -0.32, p = 0.07$. Participants reporting that they were stressed on their self-reports tended to have a higher PSS score (i.e., higher perceived stress), and had a lower threshold of being predicted as



Each row represents a feature; each column represents the comparison of stressed and not stressed groups. Colors represent positive(red) or negative(blue) differences. For example, the heart rate (HR) is significantly higher (i.e., red) during social stressor compared to during video-based resting. (S: social stressor; C: cognitive stressor; P: physical stressor; ALL: all three stressors; R: video-based resting; Reported_1: stressed based self-reported response; 0: not-stressed based on self-reported response and video-based resting; $*p < 0.05, **p < 0.005$).

Figure 3.6: Differences of means for each feature between stressed and non-stressed (in both subjective and objective stress).

Table 3.6: Model performance with using the best threshold for each participant. Mean (S.D.) are reported.

Model	Precision	Recall	Specificity	F-1 score	Accuracy	AUROC
SVM[1]	.732(.014)	.639(.017)	.930(.005)	.682(.013)	.864(.105)	.745(.007)
RF[2]	.649(.024)	.620(.015)	.899(.010)	.634(.015)	.831(.091)	.768(.003)
AdaBoost	.741(.044)	.422(.029)	.955(.012)	.536(.025)	.837(.110)	.721(.016)
GBDT[3]	.705(.052)	.603(.021)	.923(.018)	.650(.033)	.850(.097)	.755(.014)
LR[4]	.734(.014)	.581(.017)	.937(.006)	.648(.007)	.856(.116)	.694(.024)

[1]Support vector machine; [2]Random Forest; [3]Gradient Boosting Decision Trees; [4]Logistic regression.

stressed. With a lower threshold, the machine learning models have a higher probability to predict the signals as stressed.

Taking advantage of this observation, we trained the linear regression model with ridge regression between the best threshold and the PSS score on the training set, Then, in the test set, we generate the threshold from the trained regression model for each participant. Unlike the best threshold, which needs the ground truth label in the test set to select for each participant, the regression model is obtained from the training set, and the participants in the test set do not rely on any specific ground truth labels, it can be generalized to new, unseen participants.

Table 3.7 shows the LOSO cross-validation results with the personalized threshold retrieved from the linear regression. The SVM model had the best performance with an F-1 score of 0.599. The performance of the models dropped compared to the best threshold models (Table 3.6) but were still considerably better than the fixed threshold (Table 3.5).

Table 3.7: Model performance using the threshold derived from the linear regression. Mean (S.D.) are reported.

Model	Precision	Recall	Specificity	F-1 score	Accuracy	AUROC
SVM[1]	.598(.018)	.601(.011)	.879(.007)	.599(.014)	.810(.102)	.751(.011)
RF[2]	.592(.017)	.576(.016)	.888(.009)	.583(.018)	.806(.093)	.775(.006)
AdaBoost	.332(.139)	.021(.009)	.988(.003)	.039(.017)	.783(.156)	.714(.019)
GBDT[3]	.556(.012)	.603(.011)	.856(.006)	.579(.010)	.794(.116)	.759(.012)
LR[4]	.504(.010)	.564(.012)	.833(.005)	.532(.010)	.767(.109)	.693(.025)

[1]Support vector machine; [2]Random Forest; [3]Gradient Boosting Decision Trees; [4]Logistic regression.

3.5 Discussion

The impact of stress on the development of chronic and diseases is well-acknowledged—with its label as a "silent killer." [49]. In spite of its considerable effects on long term health, we have limited mechanistic characterization of the causal underpinnings of stress. An understanding of the interplay between subjective and objective markers of stress can provide preliminary insights to address the deleterious effects of stress. There is limited consensus in the current literature regarding the associations between subjective and objective stress. For example, Föhr et al. found that subjective, self-reported stress was associated with objective HRV-based stress and recovery, but was affected by several external factors (i.e., physical activity and body composition). Others have noted that objective and subjective measures of stress, in fact, measure different things, and may have different pathogenic consequences [169, 185], although both could lead to adverse outcomes [7]. We investigated subjective stress and objective stress separately using data-driven approaches with machine learning techniques and evaluated potential associations between them.

Towards this end, we conducted a laboratory-based study collecting physiological data on a smartwatch on objective markers of stress and subjective participant responses (using self-reports) on episodes of stress. Using machine learning techniques, we compared the performance of models that used physiological signals to detect stress with the objective markers, and with the subjective participant responses using self-reports. We found that the performance of subjective stress models was lower than that of the objective models of stress; however, the use of personalized thresholds, derived from standardized scales such as the PSS improved the performance of the subjective models of stress. This study, conducted using a commercial off-the-shelf smartwatch, affords new opportunities and directions for the

study of stress in routine situations. We discuss the directions for future research on the measurement of stress using smartwatches.

First, as opposed to prior studies our study utilized only a single smartwatch as opposed to multiple body-word sensing devices [70, 151], and achieved a reasonable performance compared to similar studies [70, 94, 151]. One of the highlights of our study is that the smartwatch usage did not add additional stress to participants compared to the complex body-worn devices used in other studies [152], which allows in capturing robust, potentially unbiased signals. With the advances in wearable technology and modeling techniques, there is considerable potential for improving the performance of stress prediction models. Closely related is the fact that we used some respiration-related features from the raw PPG signal; some of these features (RIAV_FFT, RIFV_FFT) showed significant differences between periods of stress and resting and were used for the models. This suggests potential use for utilizing respiration-related features for stress prediction. Although we are currently limited to the respiration-related features available from the PPG sensors, this provides a potential direction for future research.

Second, although commercial smartwatches, such as the one we used, afford considerable capabilities for real-time, unobtrusive capabilities for physiological signal monitoring, there are challenges to effective signal processing. We created a multi-stage data processing pipeline that could be used for future studies relying on physiological signals from smartwatches. The forward-backward filter that we used can remove high-frequency noise and baseline drift without phase shifts. Given the impact of motion artifacts on smartwatch-based sensing, our multi-stage approach could be used to mitigate the noise in the smartwatch signal data. Our approach involved creating sliding windows and associated sub-windows (10s duration with 2s step size) for feature extraction and noise elimination. Within the sub-windows, the combination of the motion detector and heartbeat pattern detection can detect both motion

artifacts and poor contact of the PPG sensor, potentially guaranteeing the elimination of the noisy data. This contributes to a higher data yield, as we can preserve data after the elimination of noisy spikes. Features extracted from the windows that are free from the motion artifacts and noise are likely to introduce fewer confounders in the prediction models.

Third, our approach of utilizing inter-individual differences as an input for model prediction offers new directions for modeling stress in free-living situations. Although in our case the model performance did not improve much compared with the objective stress model, this approach affords a realistic way to the personalized stress models. Previous studies (e.g., Hovsepian, K. et al. [70]) have explored the potential for developing personalized models based on participants' training data. However, such models are dependent on a limited amount of participant training data and often do not scale to real-world applications. Additionally, such models also tend to overfit on each participant, restricting generalizability. Smets et al. conducted a large-scale stress study with wearable sensors in a free-living setting [189]; their prediction model had an F-1 score of 0.4. In contrast, our approach utilizes a generic model, relying on a standardized stress scale for each participant, with a personalized threshold. In our models, we investigated the relationship between the survey responses and the threshold. Additional variables such as age, work environment, lifestyle and behaviors could potentially be incorporated into building a more informed personalized threshold. Such an approach can provide new directions toward a precision medicine approach for stress, utilizing personalized models that adjust for physiological stress response.

Finally, although we found that the models using subjective stressors had lower performance, our methodological approach has several pragmatic uses. Much of the research on self-reports, especially those using non-standard scales, has used it as a "gold standard", relying on its ecological validity rather than using objective measurements for comparison [15, 91]. In this study, using a novel comparison using machine learning techniques, we compared the objective

and subjective (self-reported) measures of stress. Our mechanism for such a comparison relied on appropriately timed smartwatch-based self-reports [76]. Such self-reports delivered via smartwatches offer a sustainable approach to capturing subjective markers as they offer a quick and easy mechanism for participants to respond to questions. Our response rate during the laboratory-based study was 100% across all tasks (n=96). Although not reported in this paper, the overall response rate of the self-reports during the free-living phase was nearly 90% with an average response time of <30 seconds, showing the potential for participant compliance to such short self-reports. As such the smartwatch-based self-report approach can be particularly useful for capturing subjective responses in a variety of settings.

We acknowledge several limitations of this study. This was a single site study with 32 participants, and as such the results may not be generalizable. This is because the laboratory-based study is unlikely to account for the complexities associated with capturing physiological signals in free-living settings. We did not counterbalance the order of presentation of the stressor tasks. This may have affected the stress perception later, like the cold stressor tasks. It is also potentially possible that the stressor tasks did not induce the necessary stress in the participants, which may have affected their subjective self-reported responses. The stressor tasks were developed from previously used experiments using wearable sensors [70]. These tasks were simplified versions of the Trier Social Stress Test [97] and the physical stressor. Although previous literature has shown that deep learning techniques could potentially address the motion artifacts and noisy data (See e.g., [181]), the relatively small set of participants made it difficult to apply such models to our data. As such, we relied on our proposed data modeling algorithms to remove noisy fragments of the physiological signals to ensure maximum data availability.

This was an exploratory study comparing the performance of machine learning-based models for stress prediction using subjective and objective markers of stress. Although the subjective

stress models had a lower performance when compared to objective stress models, additional research with a potentially larger sample of participants is required. Our approach, however, establishes an approach for commercial smartwatch-based stress detection, developing and comparing objective and subjective stress prediction models, and a framework for personalized stress prediction models. Finally, we did not report on the data collected for a day, once the participants completed the laboratory-based portion of the study. The purpose of data collection during this phase was to evaluate the viability of collecting self-reports in the wild and mapping them to corresponding physiological signals from a smartwatch. However, because of the relatively small data sample (1 day), we could not perform any meaningful computational analyses. We are currently exploring the possibility of expanding our models for stress prediction in free-living situations.

Chapter 4

Multi-Task Learning for Randomized Controlled Trials with Wearables

In this chapter, we present the application of wearables in a randomized controlled trials with integrated intervention. We proposed a multi-task learning framework to enhance the depression outcome predictions. Depression is a long-term disease with more adverse impacts compared to the stress.

4.1 Introduction

A *randomized controlled trial (RCT)* is considered the gold standard for evaluating treatment efficacy, including in the case of mental health interventions [55]. Patients enrolled in an RCT are randomized into two groups: an *intervention group* and a *control group*. Statistical methods (e.g., survival analysis and analysis of variance) are often used to assess the differences between the two groups to determine *population-level* differences and hence the effectiveness of an intervention. Although the statistical methods are powerful in assessing the value of an

intervention for clinical practice, they do not help in assessing "which patient" can achieve the desired outcome, if treated with a particular intervention. For behavioral interventions, sometimes outcomes are achieved without any treatment (e.g., a wait-and-watch approach). As such, predicting whether an intervention can have a potential impact on a "specific" patient is of great significance, given that interventions are often expensive and require time investment by both clinicians and patients. To support personalized predictions in conjunction with RCTs, a machine learning (ML) model can be trained based on data from an RCT and be used to predict the outcomes of an individual with and without the intervention. Such a predictive model can assist a physician in determining whether a specific intervention is suitable for that patient. For example, if a patient has a high likelihood of having a positive outcome without receiving the intervention, a physician may stick to the wait-and-watch approach; if a patient is likely to have a positive outcome with the treatment, the physician may prescribe the treatment for the patient. Conversely, if the patient has a high likelihood of having a negative outcome, the physician can revise the current treatment plan accordingly. This is the essence of *precision medicine*–facilitating patient-centered decisions and personalized treatment [14].

In this paper, we exploit ML techniques for personalized predictions in the context of an RCT designed to evaluate an integrated intervention for depression. Depression is a serious mood disorder; The World Health Organization (WHO) estimates that there are over 300 million people worldwide living with depression [134]. This "silent killer" is a major public health burden costing more than $1 trillion US dollars every year [61]. Wearable devices provide a convenient way for continuous remote activity monitoring, owing to their popularity, pervasive availability and relatively low cost. Recent studies [24, 127, 208, 209] have been reasonably successful in tracking depression using wearables, showing the association between depression and physical activity [131, 164]. Although previous studies with wearables were

observational in nature, we focus on developing a predictive model using both the control and the intervention groups in an RCT.

ML models in conjunction with RCTs often employ separate models for different groups of patients. If the model is developed on the control group to predict the clinical outcome without intervention, it is called *risk modeling* [90]. If the model is developed on the intervention group to predict the intervention outcomes, it is called *treatment-specific modeling* [90]. However, separate models may not be suitable for RCTs in mental health interventions due to the limited number of patients. It is challenging to recruit mental health patients for such studies, considering the cost of intervention. For example, our RCT recruited 106 patients with depression, who were randomized in a 2:1 ratio to receive the integrated intervention (n=71) or usual care (n=35). It is challenging to develop accurate ML models based on the small sample size, and splitting the dataset between the two groups further exacerbates the challenge. Also, separate models cannot capture the commonalities of the two groups with similar patient characteristics and target outcomes.

Instead of training separate models (i.e., treatment-specific modeling or risk modeling on either group) in conjunction with RCTs, we propose a multi-task learning (MTL) approach for learning from both groups of patients. Our proposed unified multi-task model is capable of predicting depression remission outcomes of a patient with and without the treatment, respectively. The MTL approach is motivated by the commonalities across the two groups in an RCT: (1) two groups share similar statistical characteristics at the baseline of a trial [55]; (2) both groups share the same outcome, e.g., depression remission in our RCT. Our MTL approach effectively enlarges the training dataset by combining the intervention and control groups to learn a single model. This modeling approach can potentially benefit many RCTs with small patient cohorts, which are typical for mobile health trials. Furthermore, we devise a hierarchical model architecture to aggregate data from different sources and different stages

of the trial, which allows the MTL model to capture the differences between two groups in an RCT. We demonstrate the advantages of our MTL approach over single-task learning and traditional MTL approaches using an RCT involving 106 patients monitored with wearable devices. The application of MTL techniques to RCTs is novel and provides a new frontier for precision treatment on already successful, evidence-based treatment methods. Specifically, the contributions of this work are as follows.

- We propose a novel multi-task learning model in conjunction with RCTs using clinical and wearable data. Our MTL model can exploit the similarity and differences between the intervention and control groups in an RCT.
- We utilize task uncertainties to dynamically weigh the task loss during the training processes. This technique can balance the task contributions and alleviate the negative transfers among the tasks when applying MTL in the RCT.
- We apply our MTL approach in a case study of an RCT with depression intervention treatments, which demonstrates the proposed MTL model outperforms both group-specific single-task models and traditional MTL models with hand-tuned task weights.
- We identify predictive features in our model through model interpretation, which shows the contribution of wearable data to the predictions of depression remission in our RCT.

4.2 Related Work

4.2.1 Mental Health with Mobile and Wearable Devices

Mental health disorders, such as depression, anxiety and stress, usually have common attributes [40, 199], and can have an adverse impact on our daily life [39]. Modern smartphones

offer an easy and inexpensive way to monitor physiological signals and behavioral patterns, including step count, voice, semantic location, and physical activity. Many studies [10, 16, 45, 209] have investigated the association between behavioral patterns and mental health disorders, by utilizing a smartphone. For example, Wang et al. conducted an observational study using the *StudentLife* Android application to continuously assess the impact of activities on mental well-being and academic performance. Several significant correlations between the smartphone sensor data and mental health outcomes were observed in the study. As the semester progresses and the workload increases, stress appreciably rises while positive affect, sleep, conversation and activity drop off among the student cohorts [209]. In another study with bipolar patients, a combination of increased GPS position changes, erratic accelerometer movements, and increased social activity were found to be suggestive of a manic phase [45]. Voice data from smartphones can also be applied to discover potential markers of mental illnesses using machine learning and natural language processing techniques [30]. However, challenges of standardizing voice data collection with privacy concerns remains challenging in such studies.

In addition to the smartphone-based studies, wearable devices have also played a key role in assessing mental health outcomes. Compared to smartphone sensing, a wearable device can have direct contact with the skin providing increased sensing capabilities and finer-grained data. Heart rate, oxygen saturation, and sleep measurements are pervasive on modern wearable trackers. Wearable trackers have also become part of fashion statements, increasing their adoption and adherence in some mobile health studies[72]. Zhang et al. employed a wristband tracker to monitor sleep, and associated the depressive symptom severity with the sleep quality [221]. The finer-grained tracking of electrodermal activity (EDA) level and heart rate variability measured by a wristband-type sensor were reported to be strong indicators of construction workers' physical and mental health status [81]. Similarly, Kim et

al. used a wearable wristband that recorded galvanic skin response (GSR) to detect stress in drivers, with an accuracy of 85.3% [92]. Another study by Seoane et al. suggested that multi-parametric testing (including GSR, temperature, respiratory rate, and heart rate) had superior accuracy in the detection of stress than any single measurement [179].

Recently, several studies [79, 113] have investigated the multi-task or multi-kernel learning to assess individual well-being. Multi-task learning is a sub-field of machine learning in which multiple learning tasks are solved at the same time. Considering that mental disorders are usually highly interconnected [40, 199], MTL could potentially benefit different mental health prediction tasks when learning together. In [113], researchers modeled depression prediction with data from different mobile platforms as an MTL problem. The proposed MTL method provided a way to analyze sensor data from different sources for the same task goal. In [79], researchers modeled five well-being components (happiness, health, alertness, energy, and stress) with an MTL support vector machine (SVM) at the same time. This modeling technique demonstrated better performance than a single-task learning (STL) model. Nonetheless, the five components were interwoven in SVM kernels, making it impossible to differentiate the feature importance and identify important features [113].

Most studies on mental health outcomes using mobile and/or wearable devices have been observational studies on a single group of patients, which do not include an active treatment or a comparison arm. Comparison between the two groups in RCTs can help us to delineate the underlying differences brought by intervention treatment, and reveal the important factors for determining precision treatment for patients [14].

4.2.2 Personalized Predictions in Randomized Controlled Trials

RCTs are regarded as the gold standard to test the effectiveness of mental health treatment. Other than the standard statistical approaches, researchers have explored utilizing machine learning models in RCTs [23, 93, 158, 207], to determine individual-level predictions. Previous wearable studies usually belong to the risk modeling category [90] to estimate the potential risk for the user without intervention treatment. The treatment-specific modeling, on the other hand, can help in determining which patients are likely to respond (or not respond) to the treatment. Chekroud et al. [23] conducted an RCT to evaluate the efficacy of the antidepressant treatment, and built treatment-specific models with 25 predictive variables. Their model demonstrated an accuracy of 64.6% [23], considerably better than a random guess. Owing to the nature of large inter-individual differences, the accuracy in similar mental health studies usually ranges from 60% to 90% [14, 77, 143]. Lhmig et al. [75] presented an anxiety level detection study using machine learning tools on an RCT. Even though it was a four-group RCT, only three groups of data were used to build the model, and bagged trees proved to be the most suitable classifier (with an accuracy of 89.8%) in their study. In another RCT study on the use of a music-based intervention for relaxation [158], the authors developed a decision tree combining data from both groups in the RCT at the same time. The decision tree model used the group indicator to generate the leaves, which is similar to building separate models for each group. There is actually no information exchange between the groups. Also, the decision tree method tends to suffer from overfitting. Other than directly predicting the final outcomes, Wallert et al. [207] used a supervised machine learning model to predict the treatment adherence in an RCT for the intervention group, as adherence to the treatment is a key factor for the success of a positive outcome. Most clinical-related RCT studies [23, 93, 158, 207] focus on either treatment-specific modeling or risk modeling.

Adaptive trials [3, 46, 62] have also been developed to evaluate personalized interventions as well. For example, the sequential multiple assignment randomized trials (SMART) can operationalize strategies leading to individualized sequences of treatment [3, 124]. A SMART trial involves more than one randomization process during a trial, which will occur at different time points based on treatment responses, facilitating potentially improved outcomes for the patients [3]. Moreover, the N-of-1 trials have become popular in devising personalized interventions [46, 99, 108, 120], and are focused on devising optimal therapy for a single individual, via periodic switching from active treatment to placebo or between different types of active treatments [99]. A strength of the SMART and N-of-1 trials is that they are designed to personalized interventions prospectively, while our machine learning models are trained and validated on data collected during RCTs retrospectively. However, SMART and N-of-1 trials are complex and require considerably more resources and effort than regular RCTs. As such, regular RCTs remain a widely used approach to evaluating an already-developed intervention with a theoretical basis [3]. Additionally, SMART may have imbalanced stratified random allocations and the N-of-1 focuses on a single patient's intervention, making it difficult to develop generalizable inferences. In this work, we focus on novel machine learning techniques in conjunction with traditional RCTs, providing personalized predictions for an evidence-based treatment approach.

4.3 Clinical Trial and Data Processing

In this section, we describe the clinical trial, problem formulation, collected data, and data preprocessing in our study.

4.3.1 Clinical Trial

Our clinical trial (ClinicalTrials.gov, NCT #03841682) was designed to examine the patient's response to an integrated collaborative care intervention for co-morbid depression and obesity. In this paper, we will focus on the prediction of depression. We recruited a sample of 106 adults from March 2019 to March 2020 from the internal medicine clinic at an academic medical center. The participants met the following inclusion criteria:

- at least 18 years old and not pregnant,
- depression (PHQ-9 [101] scores no less than 10),
- with body mass index no fewer than 30.0 (or 27.0, if Asian),
- with no significant medical comorbidities(e.g., diabetes or cardiovascular disease),
- with no psychiatric comorbidities (e.g., psychotic or bipolar disorders).

Study coordinators obtained written informed consent from each participant. Among the patients, 77.5% were female and 22.5% were male. 18.3% were Non-Hispanic White, 57.8% were African American, 2.8% were Asian or Pacific Islander, 14.1% were Hispanic and 7.0% were of other races. The average age is 46.7 (SD=11.7). Figure 4.1 shows the diagram of our RCT study.

During the orientation sessions, patients completed baseline assessments consisting of a series of surveys and clinical measurements, then were randomized in a 2:1 ratio to receive the integrated intervention (n=71) or usual care (n=35). Each patient had been followed for 6 months (primary endpoint). The integrated intervention included behavioral activation for depression care management over 6 months. There were check-point visits at 2 months, 4

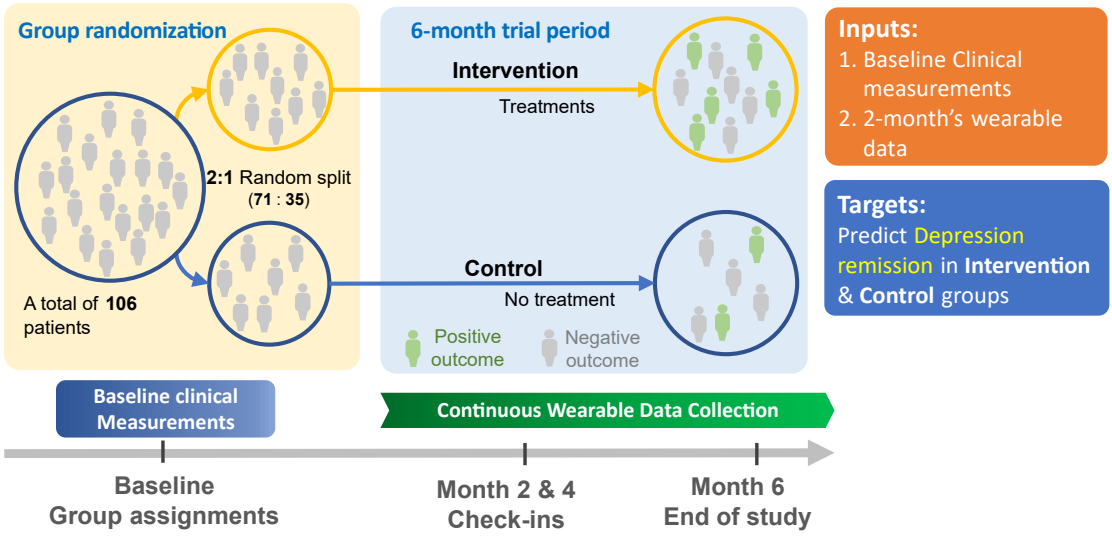


Figure 4.1: Diagram of our RCT study.

months and 6 months, during which we collected required assessments. The patients were required to continuously wear a wearable activity tracker (Fitbit Alta HR, San Francisco, CA[4]) during the entire trial. To promote retention and compliance with the trial, the study coordinators adopted an incentive strategy, which includes a reward of up to $220 [115]. The bottom part of Figure 4.1 shows the timeline of our RCT study.

We chose depression remission as our primary target, as one of the goals of the integrated intervention was to mitigate depression symptoms. Depression remission was defined as having an SCL-20 score below 0.5 [25, 26]. If a patient achieved remission, we marked that as a positive outcome. Specifically, our machine learning targets are to (1) predict the probability of one patient having depression remission with intervention treatments (in the intervention group), and (2) predict the probability of one patient having depression remission without intervention treatments (in the control group). The baseline clinical characteristics and first 2-month wearable data were used as the input since we want to have the outcome estimation at an early stage for each patient.

[4]https://www.fitbit.com/gb/shop/altahr

4.3.2 Data Collected

We excluded patients (1) who failed to take the end-trial (i.e., 6-month) assessments for the outcome labels, and (2) who had a total yield of wearable data lower than 10%. Fitbit provides minute-by-minute heart rate readings. The heart rate reading only exists when the patient correctly wears the tracker. Ideally, we would have collected 1440 points of heart rate (HR) for each patient in a day. The total yield of the wearable data was calculated as the portion of total valid heart rate readings during the study period. We selected the total yield threshold of 10% for the trade-off between patient coverage and data quality [106]. After excluding the patients who failed to satisfy the two criteria, 89 patients remained for the following analysis with 59 patients in the intervention group and 30 patients in the control group. No significant correlations were found between the patient exclusions and the group splits (χ^2 test of independence, degree of 1, $p > 0.05$). We used two major sources of data in our analysis: clinical characteristics at baseline and wearable data during the first 2-month period.

Clinical Characteristics

At baseline, we collected a variety of clinical characteristics including multiple self-reported surveys and clinical measurements. These surveys and measurements were administrated by the clinicians or professional therapists. To reduce potential confounders from other unrelated variables, we focused on the depression-related candidate predictive variables verified by the clinical experts. Table 4.1 shows the baseline clinical characteristics used in our analysis. Between the two groups, the Analysis of variance (ANOVA) test for continuous variables and the χ^2 test for categorical variables were performed. We can observe that there was little difference in the baseline clinical characteristics between the two groups.

Table 4.1: Clinical characteristics at baseline in the intervention and control group

Variable	**Intervention (n=59)**	**Control (n=30)**	p **value**
Age, year	46.86 ± 12.11	47.47 ± 12.88	0.829
Female, %	76.2%	76.6%	0.824
Race/Ethnicity, %			0.191
Non-Hispanic White	18.6%	13.3%	
Non-Hispanic Black	55.9%	50%	
Asian/Pacific Islander	3.3%	0%	
Hispanic	13.5%	33.3%	
Other Race	8.4%	3.3%	
Education, %			0.497
High school/GED or less	6.7%	16.7%	
Some college	42.3%	33.3%	
College graduate	28.8%	30%	
Post college	22%	20%	
Weight, kg	100.73 ± 15.46	99.54 ± 14.19	0.726
SBP, mmHg	122.95 ± 14.96	120.63 ± 15.96	0.5
DBP, mmHg	77.25 ± 8.69	75.83 ± 10.45	0.498
Leisure, MET mins/week [98]	654.58 ± 819.72	953.20 ± 1131.43	0.158
Work, MET mins/week [98]	336.27 ± 937.05	217.33 ± 999.95	0.581
Energy expenditure, kcal/kg/d [116]	33.52 ± 2.23	33.64 ± 2.32	0.807
SPSI-R:S raw score [31]	12.88 ± 2.41	13.13 ± 2.40	0.639
PPO raw score [31]	10.59 ± 4.54	12.58 ± 3.33	0.037
NPO raw score [31]	7.07 ± 3.48	8.62 ± 3.52	0.051
RPS raw score [31]	10.46 ± 4.48	11.29 ± 3.67	0.382
ICS raw score [31]	4.05 ± 3.65	4.74 ± 3.22	0.382
AS raw score [31]	5.54 ± 4.54	4.84 ± 3.46	0.461
PROMIS sleep disturbance t score [218]	57.96 ± 7.44	57.08 ± 8.05	0.611
PROMIS sleep impairment t score [218]	56.95 ± 8.80	54.83 ± 9.09	0.291
SCL-20 score [47]	1.21 ± 0.67	1.15 ± 0.59	0.695
GAD-7 score [195]	7.10 ± 5.12	6.63 ± 3.77	0.659
PTSD severity score [5]	36.12 ± 14.57	33.77 ± 10.25	0.432
SF-8 physical component score [201]	45.00 ± 8.48	47.38 ± 8.30	0.21
SF-8 mental component score [201]	39.65 ± 11.37	42.57 ± 9.00	0.225
COPE total scores [19]	57.02 ± 12.09	57.40 ± 14.63	0.896
BRISC total scores [215]	30.12 ± 6.23	31.07 ± 6.06	0.495

Abbreviations: AS, avoidance style; COPE, COPE Inventory survey, including 14 components [19]; BRISC, BRISC questionnaire of emotional resilience and self-efficacy Survey, including 3 components [215]; DBP, diastolic blood pressure; GAD-7, generalized anxiety disorder scale-7; ICS, impulsivity/carelessness style; MET, metabolic equivalent task; NPO, negative problem orientation; PPO, positive problem orientation; PROMIS, Patient-Reported Outcomes Measurement Information System; PTSD, post-traumatic stress disorder; RPS, rational problem solving; SBP, systolic blood pressure; SCL-20, Symptom Checklist-20; SF-8, Short Form 8 Health Survey; SPSI-R:S, Social Problem Solving Inventor -Revised: Short Form.

Wearable Data

Fitbit activity trackers provide a variety of data, including the minute-by-minute heart rate, energy consumption, sleep and other activity measurements. Considering our study spanned several months, we analyzed the wearable data at the day-level. The fine-grained measurements from wearable devices were aggregated into daily semantic features. Specifically, the following daily semantic features were extracted to depict the patients' activity characteristics.

- **Sedentary minutes (sedentaryMinutes).** Fitbit provides estimations of the active minutes through the metabolic equivalents (METs) [82, 214]. One MET is the rate of energy during rest or sitting quietly. This feature represents the duration when a user's MET is less than or equal to 1 during a day.
- **Lightly active minutes (lightlyActiveMinutes).** Analogous to the sedentary minutes, the lightly active minutes correspond to the duration with METs greater than 1 but less than 3. This feature represents the duration when a user is in a lightly active state during a day.
- **Minutes of heart rate zone in fat-burn (HRzoneFatBurnMinutes).** The minutes of heart rate zone in fat-burn is measured based on the heart rate sensors and ages. In decreasing order of intensity, Fitbit defines four zones of heart rate: peak, cardio, fat-burn, and out-of-range. When working out in the fat-burn heart rate zone, our body consumes energy from the fat stores [18]. And this feature represents the duration when a user's heart rate is in the fat-burn zone during a day.
- **Minutes of heart rate zone in cardio (HRzoneCardioMinutes).** Similar to the minutes of the heart rate zone in fat-burn, this feature represents the duration when the heart rate is in the cardio zone during a day.

- **Total walking distance (distanceTotal).** The total walking distance is calculated based on the step counts and/or GPS locations. This feature represents the total amount of distance traveled during a day.

- **Activity calories (activityCalories).** Fitbit combines the basal metabolic rate (BMR) [63] and the activity data to estimate the calories burned. This feature represents the total calories the user consumes during a day.

- **Minutes awake in main sleep (minutesAwake).** Fitbit captures the stages of sleep based on the motions and heart rate. The main sleep is the longest sleep of the day, which is usually overnight. There could be multiple records of sleep. In our study, we only considered the main sleep. This feature represents the duration when a user is awake in the main sleep.

- **Restless count in main sleep (restlessCounts).** Analogous to the minutes awake in main sleep, Fitbit captures the restless counts using the motions and heart rate sensors. This feature represents the restless count in the main sleep.

- **Efficiency in main sleep (sleepEfficiency).** Fitbit provides the calculation of the efficiency of sleep [51]. We directly adopted the calculated efficiency from the device as a daily feature.

- **Time in bed of main sleep (timeInBed).** This feature represents the duration spent in bed during the main sleep, including all sleep stages.

4.3.3 Wearable Data Preprocessing

The 10 daily semantic features from the wearable data have been shown to be effective in similar studies on mental health outcomes [113, 208]. For each patient, we used these 10

features for 60 days (i.e., a total of 60×10 wearable data points during the first two months). Even though the daily semantic features aggregated the minute-by-minute data to distill the information, 600-dimension data frame is too large to build a valid machine learning model, due to the "curse of dimensionality" [174]. As such, we employed a high-level feature engineering approach [106] to lower the input dimensions of wearable data. We applied the Singular Spectrum Analysis (SSA) to each of the daily semantic features. SSA is a nonparametric spectral estimation method for time series data, which can decompose the time series into a sum of components. Using the first component from the SSA, we can denoise and impute the time series data. Then, five statistical features (i.e., maximum, minimum, median, slope, and intercept) were extracted from the first component of each daily semantic feature time series. The slope and intercept were obtained via a linear fit of the first component. As a result, we transformed the original wearable data into 5×10 high-level statistical features. These 50 statistical features were flattened as machine learning model input candidates.

4.3.4 Feature Selection

Even though the high-level wearable features reduced the input dimension, there remained 108 features (consisting of 58 clinical characteristic features and 50 wearable high-level statistical features). Considering that there were only 89 patients in our analysis, we employed an additional univariate feature selection [38] on the training dataset in our pipeline. Univariate feature selection works by selecting the top features based on statistical tests, which can effectively reduce the input dimensionality while improving the generalizability of machine learning models [38, 57, 172]. First, we performed statistical tests in the combined group between the patients that have a positive outcome (i.e., depression remission) and the patients that have a negative outcome (i.e., no depression remission). For continuous variables, ANOVA tests were applied, and for categorical variables, χ^2 tests were applied. Then, we ranked

the clinical characteristics and wearable features based on the $p-$value from the statistical analysis, respectively. The top 10 features from each feature category were selected as the machine learning inputs [103].

4.4 Multi-task Learning for Randomized Controlled Trials

In this section, we elaborate on our proposed multi-task learning model. There are two primary tasks in our study: (1) treatment-specific modeling: to predict if a patient achieves depression remission in the intervention group, and (2) risk modeling: predict if a patient achieves depression remission in the control group. To simplify, we used the intervention task and the control task to represent the prediction tasks in the corresponding group, respectively. We proposed a multi-task learning (MTL) framework to learn the two tasks simultaneously. The MTL is inspired by human learning activities where people often apply the knowledge learned from one task to help learn another task. We exploited the commonalities of the two tasks, improving one task's performance by knowledge transfer from the other task. Unlike the previous MTL on mental health studies that focus on outcome transfers (e.g., mood and stress) [79], our MTL model focuses on group transfers. The rationale behind the group transfer is that the two tasks corresponding to the groups have the same prediction target (i.e., having depression remission or not), and patients in the intervention group and control group have no statistical difference at the baseline. However, there are still several challenges for building MTL models for group transfer in RCTs:

- **Non-unified data**– In an RCT, a patient can only have the outcome in either the intervention or the control group, which means we cannot have the two task labels for one patient at the same time. Traditional MTL models usually have all the task labels

for each data sample. Besides, the available features may also be different in the two groups (e.g., extra treatment measurements). To train a single MTL model, we need to handle the non-unified data between the intervention and control groups.

- **Task weight optimization**– MTL learning needs to assign task weights during the training, which controls task contributions to the whole model. The negative transfer may occur when the task weights are not optimal, thus degrading the overall performance. We need to find an optimal way to assign the task weights.
- **Limited Dataset size**– Even though the MTL can enlarge the training dataset by combining the two groups of patients during training, the total number of patients could still be limited. We need to avoid potential overfitting due to small sample size.

In the following subsections, we describe our MTL framework that addresses the above challenges. While we present the MTL framework in the context of the clinical trial described above, the approach can be generalized to other RCTs. To our best knowledge, our work is the first MTL framework specifically designed for RCTs.

4.4.1 Multi-task Learning Model Architecture

To exploit the commonalities as well as the differences in the two groups from an RCT, we proposed a two-layer MTL framework with hierarchical inputs, as shown in 4.2a. The hierarchical inputs accommodate the discrepancies of the inputs. In an RCT, the data collected after randomization may show differences between the intervention and control groups, due to treatment effects. In order to learn the unique characteristics after the group randomization, we feed the data before the randomization into the shared layer and the data after the randomization into the task-specific layer. This framework can be easily adapted to other RCTs.

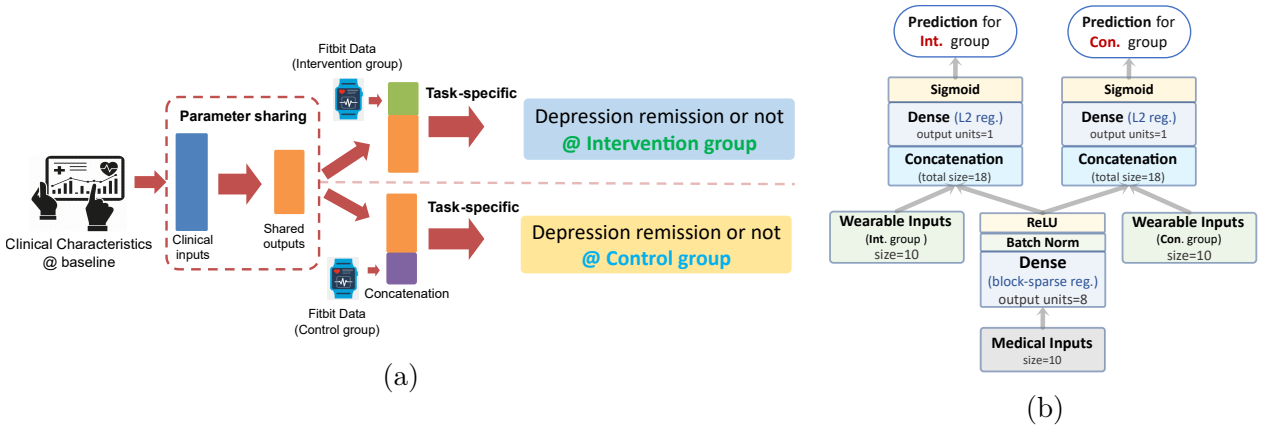


Figure 4.2: (a) MTL framework for randomized controlled trials; (b) MTL Model structure.

Given the small sample size, we only employed two dense layers for our MTL model, limiting model complexity. In the first layer, we feed the baseline clinical characteristics and force a hard parameter sharing at the first layer. Clinical characteristics at baseline showed no statistical difference between the two groups, so we wanted to capture the commonalities between the groups. The first shared layer has an output size of 8, followed by a batch normalization and a ReLU activation function. Since the wearable data were collected after the start of the treatment intervention, we do not feed the wearable features into the first layer. Instead, we concatenate the wearable features to the output of the first layer, and use the concatenation as the input for the second layer, a task-specific layer. There is no parameter sharing between the tasks in the second layer, for accommodating the differences between the groups. The final prediction for each task is from the corresponding task-specific layer followed by a sigmoid function. Figure 4.2b illustrates the model architecture.

To overcome potential overfitting and enforce parameter sharing, we employed the two types of regularization in our MTL model. The first one is the block-sparse regularization [78, 159], which is to enforce the sparsity of the parameter matrix in the first layer. We assume that two tasks in the RCT share a set of features, and other non-shared features should have

small or zero weights. Let matrix $A = \{a_1, a_2, ..., a_d\} \in \mathbb{R}^{k \times d}$ be the parameters of the shared layer, where a_i is a column vector, d is the number of input features at baseline and k is the dimension of output from the shared layer. Usually, we have k smaller than d to ensure dimension compression and parameter sharing. Based on the block-sparsity assumption [78], matrix A should only have a few columns with non-zero weights for the shared features, and other columns have zero weights for the non-shared features. We employed the mixed-norm constraints [159], which can enforce the block-sparsity for A. Basically, it first applies an l_2 norm on the feature column vector a_i, then applies an l_1 norm:

$$R_{blk} = \| |a_1|_2 , |a_2|_2 , ..., |a_n|_2 \|_1 \tag{4.1}$$

Next, we applied the second regularization to the task-specific layers. Similar to the ridge regression, we add the l_2 regularization to the loss function, which can lower the complexity of the models. Let column vector $b_j \in \mathbb{R}^{1 \times m_j}$ be the task-specific layer parameters for the jth task, where m_j is the number of the input dimension of the task-specific layer for the jth task. We have the task-specific regularization:

$$R_{tsk,j} = \|b_j\|_2 \tag{4.2}$$

4.4.2 Training MTL with a Non-unified Dataset

Another challenge of applying MTL in RCTs is to handle the non-unified labels of the dataset. Traditional MTL models usually assure the one-to-many structure; that is, a single training sample has all the task labels at the same time. However, each sample in our case only has one valid task label. It is the label either in the intervention group or in the control group. As such, we adopted a label mask for each sample during training. When calculating the loss, we only kept the task output corresponding to the sample's group, and ignored the output from the other task. For example, if a patient is from the intervention group, we only calculate the patient's loss for the intervention task. Given that we used the batch training process and a single batch consisted of the samples from both groups, the overall loss contained the information from both groups. The outcome in our study is binary for each task(i.e., depression remission or not), so we adopted the binary cross-entropy loss. Formula 4.3 shows the overall masked classification loss function for all tasks:

$$Loss_{clf} = \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} w_t \cdot mask_{n,t} \cdot BCE(y_{n,t}, \hat{y}_{n,t}) \tag{4.3}$$

where w_t is the weight for the task t, T is the total number of tasks, N is the total number of training samples, BCE is the vanilla binary cross-entropy loss, $y_{n,t}$ is the output of the n-th sample for task t, and $\hat{y}_{n,t}$ is the ground-truth label of the n-th sample for task t. The $mask_{n,t}$ is the mask for each sample ($mask_{n,t} = 1$ only if the n-th sample belongs to task t). For example, if the sample is from the intervention group, the mask will be 1 for the intervention task and 0 for the control task. The MTL model always has all task outputs for

each sample, but we only have one ground-truth label from one task. The masks enable us to calculate the loss only from the tasks with corresponding labels. The summation of the masked losses from all samples enables the MTL model to utilize the information from both groups.

4.4.3 Dynamic Task Weights

In the above subsection, we elaborated on the masked classification loss. There is a task weight w_t in the formula, which controls the contribution of each task to the overall model. The weights of the tasks need to be carefully tuned; otherwise, negative transfer may occur between tasks. For instance, if we put a large weight on one task, the model could ignore the information from the other task, incurring a performance drop for the task with a smaller weight. Manually tuning w_t could be time-consuming, and it is often hard to achieve the best performance for every task in a single MTL model. So, we employed a dynamic weight tuning technique in our MTL framework, which has demonstrated good performance in computer vision problems [89]. Previous work [89] focused on a unified dataset. Each sample has all task labels at the same time. The core idea of the dynamic weights is to use the task uncertainty to weigh the loss for each task during the training. Large uncertainty of the task means there could be a large error for the task, so we want to lower its contributions to the overall MTL model. A task output can be regarded as an object's state based on some observations. In physics, an object's state can be modeled with the Boltzmann distribution [171], and the object is less stable when it is at a higher temperature. The uncertainty of a task is akin to the temperature of an object in the Boltzmann distribution. So, we adapted the task probability output with an uncertainty parameter. The uncertainties are trainable parameters in the loss function of the MTL model, which can be dynamically updated during the training process to adjust the task weights. In our study, we have two binary classification

tasks with the non-unified dataset. We extended the dynamical task weighing to the MTL in conjunction with RCTs, as illustrated below.

For a binary classification problem, we have two outcomes: positive and negative. We adopted the sigmoid function for the probabilities of positive outcome p_t, where t means the t-th task. Therefore, the probability of a negative outcome is $1 - p_t$. We added a trainable uncertainty factor σ_t to the sigmoid function to mimic the temperature in the Boltzmann distribution. The σ_t was automatically adjusted in each batch training:

$$y'_{n,t} = p(y_{n,t} = 1 | f_t(x_n), \sigma_t) = Sigmoid(\frac{1}{\sigma_t^2} f_t(x_n)) = \frac{e^{f_t(x_n)/\sigma_t^2}}{1 + e^{f_t(x_n)/\sigma_t^2}} \tag{4.4}$$

where $y'_{n,t}$ is the updated probability output of the n-th sample for task t with the uncertainty parameter σ_t^2, $f_t(x_n)$ is the output of the n-th sample before the sigmoid activation function for the task t. Basically, if we have large σ_t^2, there could be larger uncertainty for the task, thus a lower probability output. σ_t is squared to match the form of standard deviation in the normal distribution. Ignoring w_t and using the probability output with the uncertainty in

Formula 4.3, we have:

$$Loss'_{clf} = \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} mask_{n,t} \cdot BCE(y'_{n,t}, \hat{y}_{n,t}) \tag{4.5}$$

$$= \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} mask_{n,t} \cdot [-\hat{y}_{n,t} \cdot log(y'_{n,t}) - (1-\hat{y}_{n,t}) \cdot log(1-y'_{n,t})] \tag{4.6}$$

$$= \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} mask_{n,t} \cdot [-\hat{y}_{n,t}\frac{f_t(x_n)}{\sigma_t^2} + log(1+e^{f_t(x_n)/\sigma_t^2})] \tag{4.7}$$

$$= \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} mask_{n,t} \cdot [-\frac{1}{\sigma_t^2}\hat{y}_{n,t}f_t(x_n) + \frac{1}{\sigma_t^2}log(1+e^{f_t(x_n)}) + log\frac{1+e^{f_t(x_n)/\sigma_t^2}}{(1+e^{f_t(x_n)})^{\frac{1}{\sigma_t^2}}}] \tag{4.8}$$

$$= \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} mask_{n,t} \cdot [\frac{1}{\sigma_t^2}BCE(y_{n,t}, \hat{y}_{n,t}) + log\frac{1+e^{f_t(x_n)/\sigma_t^2}}{(1+e^{f_t(x_n)})^{\frac{1}{\sigma_t^2}}}] \tag{4.9}$$

$$\approx \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} \frac{1}{\sigma_t^2} \cdot mask_{n,t} \cdot BCE(y_{n,t}, \hat{y}_{n,t}) + \sum_{t=1}^{T}\log(\sigma_t) \tag{4.10}$$

In Equation (4.9), we used an approximation: $(1+e^{f_t(x_n)})^{1/\sigma_t^2} \approx 1/\sigma_t^2(1+e^{f_t(x)/\sigma_t^2})$ [89]. When $\sigma_t = 1$, it becomes equality. In Equation (4.10), We can see that $1/\sigma_t^2$ happened to be at the same place as the original w_t, which is to control the weights for each task in the loss. If we have large uncertainty σ_t, $1/\sigma_t^2$ will be small, thus lowering the contributions from task t. There is another term: $\log(\sigma_t)$, which can be viewed as a regularization to avoid the uncertainty σ_t to be infinity, as infinity will make the first term in the loss to be zero. Replacing the classification loss Equation (4.3) with Equation (4.10) and summing up the regularization terms, we can then obtain the updated total loss for our MTL framework:

$$Loss = \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} \frac{1}{\sigma_t^2} \cdot mask_{n,t} \cdot BCE(y_{i,t}, \hat{y}_{i,t}) + \sum_{t=1}^{T}\log(\sigma_t) + R_{blk} + \sum_{t=1}^{T} R_{tsk,t} \tag{4.11}$$

By optimizing the above loss function, we can train our MTL model for our RCT, and output predictions for both intervention and control groups simultaneously.

4.5 Evaluations

In this section, we present the detailed evaluation of our MTL model. The performances of the two tasks in the RCT were compared to each STL model. Since one of our motivations is to validate the feasibility of wearable devices in mental health studies, we also demonstrated the contribution of the wearable data in the model performances. Finally, we explored shedding more light on our MTL model with state-of-the-art deep learning explanation tools, as it is essential for applications in healthcare.

4.5.1 Evaluation Settings

Our targets are to predict: (1) the probability of one patient having depression remission with intervention treatments (in the intervention group), and (2) the probability of one patient having depression remission without intervention treatments (in the control group). We compared each prediction task between the MTL model to the STL models, respectively. The five-fold cross-validation (CV) approach was adopted to evaluate the model performances. In the CV, we stratified the whole dataset into five folds. Each fold contains the same portions of patients from the two groups. Every time we chose 4 folds to train the models, and used the remaining fold to evaluate the model performance. This procedure was repeated five times until all the folds had been used as testing once. To avoid opportune splits of the dataset, we conducted 20 runs of the CV to report the average and standard deviation of model performances.

It is worth noting that our dataset is imbalanced. 42% of the patients in the intervention group had a positive outcome (i.e., depression remission), whereas only 23% of the patients in the control group had a positive outcome. Therefore, we used the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) as the major metrics. Those two metrics can well gauge the performance with the imbalanced dataset when the positive outcomes are of more interest. We included the following three groups of models in our evaluations:

- **STL-separate**: baseline shallow STL models that were trained for the intervention and control group tasks separately.
- **STL-unified**: baseline shallow STL models that were trained on the combined groups, treating the two tasks as a single task.
- **MTL**: three MTL models, of which the only difference was the task weight assignment. MTL-1 is the MTL model that was trained only on a single group, via setting the task weight of one group at 1 and the other group at 0; MTL-fixed is the MTL model that was trained on the combined group of patients, and used a grid search to find the optimal performance of a single task; MTL-dynamic is the MTL model that was trained on the combined group of patients with the dynamic task weights.

For both STL-separate and STL-unified, we included six models: (1) support vector machine (SVM with rbf kernel), (2) random forest (RF), (3) Adaboost trees (Ada), (4) gradient boosting decision trees (GBDT), (5) logistic regression (LR) and (6) 3-layer artificial neural network (ANN). The ANN model has two hidden layers with ReLU activation, and an output layer with sigmoid activation. The STL-separate models were trained on the intervention group and control group separately, resulting in two STL-separate models corresponding to

each task. The STL-unified models were trained on the combined group of patients, and we only had one model for the two groups. It simply treats the prediction in the two groups as a single task by adding a group indicator in the input. The patients were differentiated by the group indicator, and the performances of the STL-unified models were evaluated for the two groups separately. There are three MTL models (i.e., MTL-1, MTL-fixed, MTL-dynamic) sharing the same architecture in our evaluations. MTL-1 is actually a single-task model by setting one task weight to zero. The loss function only contains the task loss from one group, so it reduces to the single-task learning. Similar to STL-separate, we trained two MTL-single models–one for the intervention group and one for the control group. This is to have fair comparisons between the MTL and STL by eliminating the impacts of model architecture change. MTL-fixed is the MTL model trained on the combined groups of patients, using the fixed task weights. We also trained two MTL-fixed models for the two groups separately. Each MTL-fixed model was optimized to achieve the best performance for only one task via a grid search to find the best task weights. MTL-dynamic is our proposed MTL that learns the two tasks simultaneously with the dynamic task weights. We trained only one MTL-dynamic model to predict the probabilities for the two groups. The MTL models were implemented with the Tensorflow framework [118]. We used an Adam optimizer (learning_rate=0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) [95], with a single batch of all training data. The number of training epochs (300) of the MTL models was empirically selected based on the train/test loss from several history runs. The hyperparameters of the STL and MTL models were tuned within the training dataset using grid-search CV [139] to achieve the best AUROC. Table 4.2 lists all the hyperparameters tuned in our experiment.

Table 4.2: List of hyperparameters for grid search CV

Model	Hyperparameter	Candidate values	Model	Hyperparameter	Candidate values
LR	penalty	"l1", "l2"	Ada	base estimator	decision stump [138]
	C	0, 0.5, 1, 10		number of trees	20, 50, 100
SVM	kernel	"rbf"		learning rate	0.1, 0.5, 1
	C	0.1, 1, 10	ANN	hidden size (1st)	16, 8, 4
	gamma	0.01, 0.1, 1		hidden size (2nd)	16, 8, 4
RF	number of trees	50, 100, 200	GBDT	number of trees	50, 100, 200
	max depth	3, 10, None		learning rate	0.01, 0.1, 0.5
	max features	5, 10, 20		max depth	3, 10, None
MTL	block regularization	1e-4, 1e-3, 1e-1			
	task regularization	1e-4, 1e-3, 1e-1			
	task weights (only for MTL-fixed)	0, 0.05, 0.1,...,1			

4.5.2 Selected Features

The univariate feature selections were performed on the training dataset. In each cross-validation (CV) split, we selected the top 10 features with the smallest p-values. Since there were variances across different splits, the selected features also varied slightly. Nonetheless, the selected features remained the same in most of the splits. As we conducted 20 runs of the five-fold CV, the maximum selected time of a feature is 100. Table 4.3 shows all the selected features and their corresponding selected times.

4.5.3 MTL vs. STL

Table 4.4 shows the performance evaluations of all the STL and MTL models. We use the group name to represent the prediction task in that group: "Intervention" is the model performance of predicting task in the intervention group, whereas "Control" is the model performance of predicting task in the control group. We first compared the STL-separate models. Each shallow model was trained separately for the two group tasks. The logistic regression (i.e., STL-LR) shows the best performances in both tasks. For STL-unified models, logistic regression also shows the best performance in the control task, whereas GBDT shows

Table 4.3: Features from univariate feature selection

Selected clinical features				Selected wearable features			
Features	**Ave. stat**	**Ave. p**	**Times**	**Features**	**Ave. stat**	**Ave. p**	**Times**
PROMIS sleep disturbance t score	15.68	2.95E-04	100	distanceTotal_median	5.54	3.55E-02	99
PROMIS sleep impairment t score	12.43	1.24E-03	100	restlessCounts_intercept	2.61	1.23E-01	99
SCL20 baseline score	8.88	5.81E-03	100	restlessCounts_min	2.75	1.19E-01	98
NPO raw score	5.41	3.03E-02	100	minutesAwake_intercept	2.31	1.45E-01	97
COPE acceptance score	5.20	3.36E-02	99	minutesAwake_min	2.94	1.17E-01	91
PTSD severity score	4.70	4.49E-02	99	distanceTotal_intercept	2.80	1.30E-01	86
Work, MET mins/week	5.06	3.10E-02	97	restlessCounts_slope	3.47	9.76E-02	84
Sex	4.27	4.77E-02	90	lightlyActiveMinutes_min	2.11	1.78E-01	75
COPE denial score	4.36	4.89E-02	89	minutesAwake_slope	2.87	1.27E-01	64
BRISC skill scores	4.16	5.21E-02	65	HRzoneFatBurnMinutes_min	1.92	1.90E-01	57
SF-8 mental component score	3.81	6.33E-02	40	restlessCounts_max	2.16	1.70E-01	42
SF-8 physical component score	3.94	5.84E-02	12	distanceTotal_max	2.00	1.95E-01	31
GAD-7 score	3.70	6.52E-02	4	HRzoneFatBurnMinutes_slope	2.37	1.72E-01	30
DBP	3.09	8.34E-02	2	sedentaryMinutes_median	1.98	1.84E-01	22
COPE plan score	3.04	8.57E-02	1	sedentaryMinutes_min	1.76	2.13E-01	12
COPE active score	2.22	1.41E-01	1	HRzoneFatBurnMinutes_median	2.56	1.69E-01	4
Energy expenditure	2.15	1.47E-01	1	lightlyActiveMinutes_intercept	1.98	1.79E-01	3
				sedentaryMinutes_slope	1.68	2.12E-01	2
				minutesAwake_max	1.21	2.77E-01	2
				minutesAwake_median	1.49	2.27E-01	1
				lightlyActiveMinutes_max	1.28	2.61E-01	1

*Ave. stat: average statistic of either ANOVA test or χ^2 test from different CV splits.
*Ave. p: average $p-$value of either ANOVA test or χ^2 test from different CV splits.

the best performance in the intervention task. However, when comparing the STL-unified with STL-separate with the same shallow models, we can find that the STL-unified usually has worse performance. These results indicate that simply combining the group of patients does not improve the performance. The shallow models are not capable to well exploit the group commonalities and differences via the group indicator.

When comparing our MTL models with the best STL models, we observe that our MTL-1 models show comparable performance to the best STL models. The MTL-1 is effectively single-task learning, since it only utilized the information from one group. There is no performance gain from the MTL-1 model, suggesting that changing the model architecture from the shallow models to the proposed 2-layer MTL model has no impact on the task performances. However, when comparing MTL-1 to MTL-fixed and MTL-dynamic, we can observe that both MTL-fixed and MTL-dynamic outperform the MTL-1 (Wilcoxon rank-sum test [41], $p < 0.05$). The MTL-fixed and MTL-dynamic models were trained on the combined

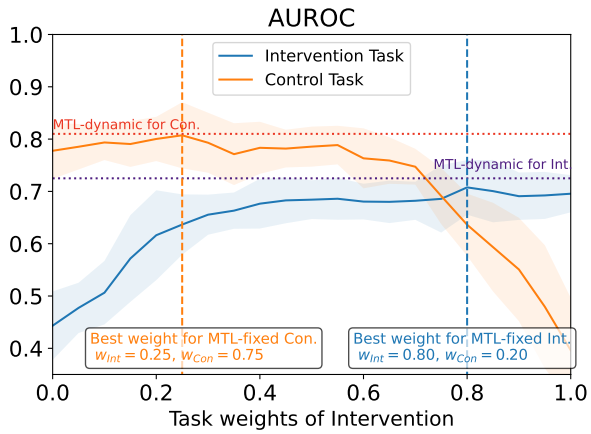
Table 4.4: Model Performance in different groups

		Intervention		Control	
Category	Model	AUROC	AUPRC	AUROC	AUPRC
STL-separate	SVM	0.607(0.053)	0.556(0.052)	0.734(0.063)	0.561(0.112)
	RF	0.667(0.040)	0.601(0.046)	0.755(0.066)	0.538(0.089)
	Ada	0.615(0.042)	0.567(0.045)	0.681(0.076)	0.433(0.086)
	GBDT	0.657(0.061)	0.582(0.060)	0.724(0.098)	0.487(0.072)
	ANN	0.659(0.040)	0.581(0.051)	0.754(0.081)	0.548(0.136)
	LR	**0.697(0.050)**	**0.636(0.063)**	**0.794(0.067)**	**0.601(0.093)**
STL-unified	SVM	0.533(0.062)	0.477(0.057)	0.683(0.107)	0.458(0.091)
	RF	0.649(0.053)	0.587(0.063)	0.704(0.077)	0.474(0.086)
	Ada	0.609(0.056)	0.562(0.062)	0.619(0.093)	0.340(0.064)
	GBDT	**0.650(0.060)**	**0.596(0.067)**	0.555(0.089)	0.332(0.076)
	ANN	0.629(0.071)	0.561(0.054)	0.734(0.045)	0.528(0.056)
	LR	0.639(0.049)	0.571(0.059)	**0.759(0.047)**	**0.569(0.084)**
MTL	MTL-1[1]	0.695(0.032)	0.641(0.051)	0.784(0.049)	0.589(0.073)
	MTL-fixed[2]	0.707(0.052)	0.653(0.046)	0.807(0.063)	0.615(0.083)
	MTL-dynamic	**0.725(0.059)**	**0.668(0.068)**	**0.813(0.077)**	**0.637(0.061)**

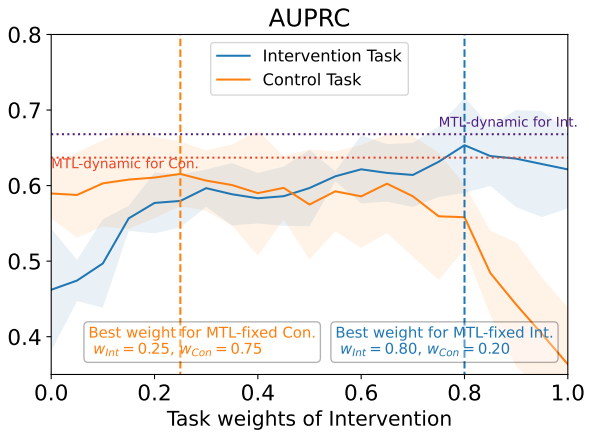
[1]MTL-1 is the single task learning but with the same architecture of our proposed MTL model.
[2]MTL-fixed is trained on the combined group, but the performances of two tasks are from separate models.

group of patients, which utilized the information from both groups. The performance gain demonstrates that it is the positive knowledge transfers between the groups that improve the performance compared to the STL.

It is also worth noting that the MTL-fixed models were trained for the task separately, even though they utilized the information from the two groups. Each MTL-fixed model was optimized for one task, using a grid search to find the weights with the best AUROC for that task. Figure 4.3 shows the performances of the MTL-fixed models with varying task weights. Without losing generalizability, we fixed the sum of task weights at 1. The x-axis represents the weight of the intervention task (w_{Int}), so the weight of the control task is $w_{Con} = 1 - w_{Int}$. From left to right, the intervention task weight increases while the control task weight decreases. When $w_{Int} = 0.85$ and $w_{Con} = 0.15$ (marked by the blue vertical line), the intervention task achieves the best performance. When $w_{Int} = 0.20$ and $w_{Con} = 0.80$ (marked



(a) AUROC for the two tasks



(b) AUPRC for the two tasks

Figure 4.3: Performances with varying task weights.

by the orange vertical line), the control task achieves the best performance. It is obvious that the two tasks achieve their best performance with different weight settings. So, we need two MTL-fixed models to achieve optimal performances for both tasks separately.

We also marked the performances from the MTL-dynamic models in Figure 4.3 with horizontal dashed lines. We can see that our MTL-dynamic models can achieve better or comparable performance than the MTL-fixed models. The difference between the MTL-dynamic model and the MTL-fixed model is the way we assign weights in the classification loss. The task weights in the MTL-fixed model were fixed and never changed during the training process, whereas the MTL-dynamic model utilized a trainable parameter–"uncertainty", to mimic the weights between the tasks. The MTL-dynamic model only needs to train once for the two tasks. Since "pseudo task weights" $(1/\sigma_t^2)$ in MTL-dynamic were dynamically updated every epoch, it is possible that we can achieve the best performances for both tasks as the training progresses [89].

4.5.4 Contribution of the Wearable Data

The wearable device played an important role in our RCT study. It fills the gap of remote monitoring with continuous data collection outside the hospital. To quantify the contributions from the wearable device, we evaluated the model performance without the wearable data. Table 4.5 shows the model performances with and without wearable data. We only demonstrate the performance of our proposed MTL-dynamic model, as it shows superior results in previous evaluations. The model with wearable data significantly outperforms the model without wearable data in both intervention and control tasks (Wilcoxon rank-sum test, $p < 0.05$), attesting the wearable data indeed encodes some information that can improve the model performances. Previous studies [113, 208] also demonstrated that wearable data show predictive power in mental health applications.

Table 4.5: Performance comparison with and without wearable data

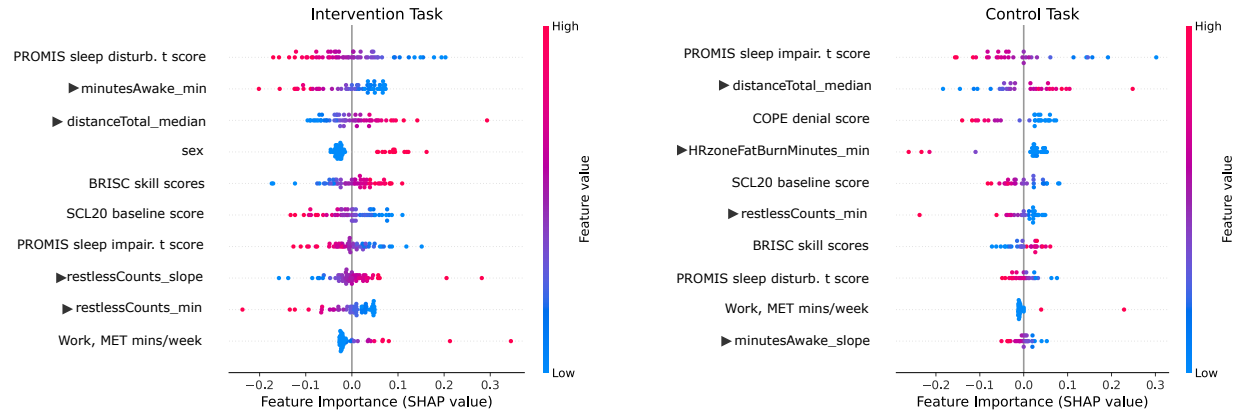
	Intervention		Control	
	AUROC	AUPRC	AUROC	AUPRC
W/ wearable data	**0.725(0.059)**	**0.668(0.068)**	**0.813(0.077)**	**0.637(0.061)**
W/O wearable data	0.652(0.041)	0.601(0.056)	0.727(0.081)	0.513(0.088)

4.5.5 Model Explanation

It is of great importance to understand the underlying logic of the model predictions, especially when a machine learning model is applied in real clinical practices. We employed the state-of-the-art model explanation tool, the SHapley Additive exPlanations (SHAP) [114], to have the model-agnostic explanations. The general principle behind SHAP is to employ the game theoretic approach to explain the output of any machine learning model with Shapley values [114]. The SHAP for deep models is built on a connection with DeepLIFT [184], using

a distribution of background samples and Shapley equations to linearize the components in the deep network. Figure 4.4 shows the SHAP summary plots for the two tasks in our MTL-dynamic model. The model was retrained on the whole dataset, with the top 10 most frequent features selected from each feature category (i.e., clinical and wearable) and the most frequent selected hyperparameters in the cross-validation. We depicted the top 10 features based on the SHAP value, and marked the wearable features with "►" in the summary plots. For each patient, we computed the SHAP values for each feature, which are shown as dots in the summary plot. A dot in the plot encodes both the true feature value and the computed SHAP value. The true feature value is represented via the color map, in which the blue color represents a lower feature value and the red color represents a higher feature value (except for Sex, where blue signifies male). The SHAP value is represented by the x-axis value. The absolute SHAP value marks the relative importance of the corresponding feature, and a positive SHAP value (i.e., on the right side of the x-axis) means the model tends to have a positive prediction, and vice versa. All the features were ranked in their order of importance in the figure, based on the average of the absolute SHAP values from all dots. For example, the sleep disturbance score (sleep_disturb) is the most important feature for the intervention task, since its average absolute SHAP value is the largest. The model tends to predict a patient having depression remission if the patient has a lower sleep disturbance score, as a lower sleep disturbance score (i.e., blue color) corresponds to a positive SHAP value (i.e., on the right side of the x-axis).

In the two SHAP summary plots, there are six clinical characteristic features for each task, and five out of the six are shared between the two tasks, demonstrating our MTL model effectively utilized the overlapping clinical features between the tasks. For both intervention and control tasks, our model tends to predict the patient will have depression remission if the patient has a low sleep disturbance score (i.e., sleep_disturb), a low sleep impairment score



(a) Shap summary plot for the intervention task (b) Shap summary plot for the control task

Figure 4.4: Model explanation for MTL-dynamic models.

(i.e., sleep_impair), a low SCL-20 score at baseline (i.e., scl20_score), and a high BRISC skill score (i.e., brisc_skill) [215]. Both low sleep disturbance score and low sleep impairment score mean a high sleep quality, and a high sleep quality marks a lower risk of depression [42]. It matches that our model has positive Shapley values for the low sleep disturbance score and low sleep impairment score. For the baseline SCL-20 score, the higher score means a more severe depressive disorder. So, it is not strange that our model has a negative SHAP value for a high baseline SCL-20 score. The median of daily distance (i.e., distance_total_median) also plays a key role in both tasks, corresponding to the fact that our model has a positive SHAP value for the high median of daily distance. Previous literature [196] has shown that exercise could be a helpful treatment to depression, and daily walking predicts an improved depression outcome[32, 121]. Our model demonstrates a similar trend as well.

4.6 Discussions and Conclusions

In this paper, we exploited machine learning (ML) models for personalized predictions in the context of an RCT. ML with RCTs usually has separate models for different groups of patients. In contrast, we formulated the outcome prediction problem for different groups

as a multi-task (MTL) learning problem, and proposed a novel MTL model for RCTs. The MTL can predict outcomes of a patient with and without the treatment, using a single model. We proposed a hierarchical input architecture, enabling the model to take advantage of the commonality and differences between two groups in an RCT. To overcome potential negative transfers, we employed the dynamic task weighing technique, which can balance the contribution of each task in the MTL model during training.

We evaluated our MTL approach on an RCT case study that was designed to test an integrated collaborative care intervention for depression. We recruited 106 patients (2:1 randomized) longitudinally monitored with wearable devices. The MTL model was trained on the dataset that combines both groups, effectively enlarging the training dataset. Our MTL model is capable of predicting depression remission outcomes of a patient with and without the intervention. The results demonstrated that the MTL with knowledge transfers between the two groups outperforms single-task models.

Since depression is usually a long-term disorder [113], automatic estimations of the outcome could be beneficial to monitoring depression status over time, and potentially assist the doctor in devising personalized treatments. Table 4.5 and Figure 4.4 demonstrate that wearable data played an important role in our MTL model, providing additional evidence that wearable devices can be used as a powerful tool to monitor depressive disorders. In the context of precision medicine, our approach contributes to streamlining the clinical point-of-care use of an already successful intervention by considering clinical characteristics and wearable-device-based activity characteristics. This not only helps in intervention choice decisions, but also in potentially changing the frequency/dose (e.g., number of times a particular therapy) of an intervention. The application of MTL techniques to RCTs is novel and provides a new frontier for precision treatment on already successful, evidence-based treatment methods.

Limitations: We note that our MTL model is designed to work in conjunction with RCTs. It assumes that a patient's treatment path does not change after group splitting, and the model needs to be trained retrospectively on groups of patients. Our method may not be applicable to clinical trials that involve adaptive interventions. For example, the sequential multiple assignment randomized trials (SMART) and N-of-1 trials can adapt the treatments for individual participants during a trial, based on their response to an intervention.

Besides, even though we have applied multiple techniques to avoid fitting, our RCT study still has a limited sample size. More confidence in our method will be gained with more and larger RCTs.

Lastly, we did not evaluate the impact of different lengths of wearable data when building the models. The lengths of the wearable data are determined by the prediction timeline. Since we were interested in having the prediction at an early stage of the intervention, we only built and evaluated the models with two-month wearable data.

Future Work: There are some future directions to advance our work in this paper. First, we can recruit more patients, and cross-validate the model in other institutions to enhance the statistical power of our analysis. Second, we can build MTL models at different checkpoints based on previous RCT data, helping in devising personalized treatments in a finer granularity. For example, our model can be trained to estimate whether a new patient should receive the treatment or not, when we only utilize the baseline data at the first visit.

Chapter 5

Predicting Mental Disorders with Wearables: A Large Cohort Study

In the previous chapter, we introduced the application of wearables in randomized controlled trials for predicting mental health outcomes with or without treatment. The study cohorts restrict to certain patients with depressive disorders in controlled settings. In order to justify the applications of wearables in a larger population, this chapter presents the detection of mental disorders in the general public with wearables.

5.1 Introduction

Depression and anxiety are among the most prevalent mental disorders, and they are usually interconnected [40]. Patients with depression often have features of anxiety disorders, and those with anxiety disorders commonly always have depression [199]. Although those two mental disorders have drawn increasing attention due to their tremendous negative impacts on working ability and job performance [117], over 50% of patients are not recognized or

diagnosed [111]. The gold standard for assessing depression and anxiety relies on clinical visits by means of questionnaires, e.g., the 20-item Symptom Checklist Depression Scale (SCL-20) [47] and General Anxiety Disorder-7 (GAD-7) [195]. Nonetheless, attaining an appointment with mental health clinicians is not an easy process and usually demands considerable time and money, thus hindering in-time diagnosis and intervention in the general public. Both patients and healthcare providers would benefit from an automated passive detection of depression and anxiety symptoms if only minimum extensive equipment is required.

The growing adoption of wearables affords a promising way for longitudinal monitoring of a range of digital phenotypes, including physical activity, heart rate and sleep. These can be used to obtain individual health profiles and indicators of depression and anxiety symptoms [68, 209]. Leveraging data-driven approaches, such indicators with remote access from healthcare providers could help narrow the gap in the diagnosis of depressive and anxiety disorders. Recently, an increasing amount of research has explored utilizing wearables or mobile devices to detect mental disorders [94, 113, 209]. Wearables are also of benefit to the patients who need or undergo mental health treatments. People with depressive and anxiety disorders sometimes hesitate to seek help or treatment because, for example, they think they can get over the symptoms on their own, fear discussing their symptoms, or simply do not know where to find the essential assistance. Wearables can remind the patients to receive treatment, and keep track of the progress of the treatment [35].

Nonetheless, previous wearable studies on depressive and anxiety disorders were usually with a small or restricted cohort [94, 113, 209]. Multiple findings suggest that physiological and psychological responses tend to be person-dependent [37, 73]. The small cohorts may not be enough to cover all inter-individual differences. To unleash the full power of the wearables on the detection of mental health disorders for the ordinary, a large dataset with a wide spectrum is essential for data-driven approaches to capture those inter-individual differences, thus

establishing personalized risk profiles. In this work, we tackled the challenges of discovering potential patients with depressive and anxiety disorders. We presented a study on a large dataset consisting of more than 11,600 participants from the "All of Us" program [203], whom were longitudinally monitored by wearable activity trackers. To the best of our knowledge, no previous research has investigated depression and anxiety detection with wearables in such a large cohort. Specifically, our contributions are in three folds:

- We built a large dataset from the "All of Us" program consisting of 11,600 participants, and statistically analyzed the wearable data and static characteristics among the participants with and without mental disorders.
- We proposed a deep learning model combining the transformer encoder and convolutional neural network, which demonstrates superior performances over other state-of-the-art models.
- We systematically investigated the model performances in various settings, and used the modern model explanation tool to illustrate the underlying feature importance of our model.

5.2 Related Work

A growing body of literature suggests that daily physical activity and human behavior patterns are associated with mental health conditions. Recent studies have used sensing data collected from smartphones and wearables to capture those daily physical activities and human behavior patterns, thus quantifying the mental health conditions [16, 30, 45, 113, 189, 209, 217]. Wang et al. conducted an observational study among college students, and found several significant correlations between the smartphone sensor data and mental health

outcomes [209]. In another study with college students, Xu et al. used association rule mining approaches to extract contextually filtered features from passively collected, time-series mobile data, and fed those features to machine learning models to predict depression with an accuracy of 81.8% [217]. Beyond the work in identifying mental health issues on campus, Kim et al. used a wearable wristband that recorded galvanic skin response (GSR) to detect stress in drivers, with an accuracy of 85.3% [92]. Zachary et al. proposed a framework with microinteraction-based ecological momentary assessment and wearable sensors to detect the stress period on pregnant women in real-world settings. Zhang et al. employed a wristband tracker to monitor sleep, and associated the depressive symptom severity with the sleep quality [221].

Beyond the physical activity and human behavior measurements, voice data from smartphones can also be utilized to discover potential mental illnesses using machine learning techniques [30]. Asif et al. [173] proposed a weakly supervised learning framework for detecting social anxiety and depression from long audio clips. However, it remains challenging to perform privacy-preserved voice data collection and analysis. Radio signals could be utilized for depression screening as well. Shweta et al. [212] explored an approach that uses data collected from Wi-Fi infrastructure for large-scale automatic depression detection, with an F-1 score of 0.85, demonstrating comparable performance to the approaches using the sensing data from smartphones and wearables. Nonetheless, the study [212] was also restricted to university campuses, even though having a relatively large cohort.

5.3 Dataset and Statistic Analysis

Our study cohort is a part of the "All of Us" research program[5] funded by the national institute of health (NIH) in the United States. The research program targets enrolling a

[5]https://allofus.nih.gov/

diverse population to accelerate biomedical research and precision medicine [203]. Multiple mobile/wearable health techniques are incorporated into the program, not only to help manage the enrollment and retention of participants, but also to gather digital phenotypes.

Participants with any Fitbit devices (Fitbit, Inc. San Francisco[6]) can share their wearable digital phenotypes via the Fitbit Bring-Your-Own-Device (BYOD) project [2], which links participants' Fitbit accounts to the "All of Us" program. More than 11,600 participants contribute to the Fitbit dataset, which contains the Fitbit daily summary time series, intraday heart rate time series and intraday step time series.

More importantly, there is important health information associated with the Fitbit data, including professional clinical surveys, electronic health records (EHRs), and biosamples. We can directly extract the diagnoses of mental health disorders from the EHRs, which contain the formatted condition codes (e.g., ICD-10 [154], SNOMED[194]) and the timestamp when the patient was diagnosed. Those diagnoses were from hospitals or clinics with professional healthcare practitioners. In addition, the demographic and some other characteristics (e.g., family disease history) are also provided in the "All of Us" program.

In the following, we elaborate on the basic information and statistic analysis of our study dataset from the "All of Us" program, targeting the predictions of depressive and anxiety disorders.

5.3.1 Labeling and Inclusion Criteria

Depressive and anxiety disorders are interconnected [199], and are often shown as comorbidities. It is sometimes difficult to separate one from another. Besides, there are some subtypes of diagnoses (e.g., posttraumatic stress disorder [220]) that can be categorized as depressive

[6]https://healthsolutions.fitbit.com/aboutus/

and/or anxiety disorders as well. As such, we included a combination of 20 diagnoses (see Table 5.1), and treated them as the defined depressive and anxiety disorders in our analysis. Those 20 diagnoses were verified by clinical experts. Among all the participants with Fitbit data, we searched the associated EHRs and labeled the participants as positive if he/she has any one of those 20 diagnoses. Participants without any one of those 20 diagnoses were labeled as negative. For convenience, we will use "positive" or "negative" to represent the participants with or without the defined depressive and anxiety disorders, respectively.

Given that the depressive and anxiety disorders are usually long-term conditions [29, 88], we chose our primary wearable data window as 60 days when performing the analysis unless specifically stated otherwise. All the data windows should satisfy a yield threshold. The yield is defined as the ratio of days with non-zero total steps to the length of the window. The threshold can be varied, which is a trade-off between participant coverage and data quality [106]. A yield of 10% was used in our primary analysis, that is, we only kept the data window having as least 10% days with non-zero steps. Figure 5.1 illustrates our window sampling strategy. For each positive participant, we extracted one data window right before the diagnosis time. If the participant has multiple diagnoses with qualified windows, we chose the earliest window. For each negative participant, we randomly sampled one window that satisfied the yield threshold. If we could not find a valid window for a participant, this participant would be excluded. After the exclusion, there remained a total of 8,996 participants with 1,247 positive cases. Table 5.1 displays the distribution of the diagnoses of participants with a valid window.

Other than the wearable data, we incorporated some static characteristics into our analysis, as mental health disorders and activity measurements were reported to be associated with some static factors (e.g., age and gender) [156, 191]. Two criteria were enforced for the inclusion of the static characteristics. First, only common characteristics that were available

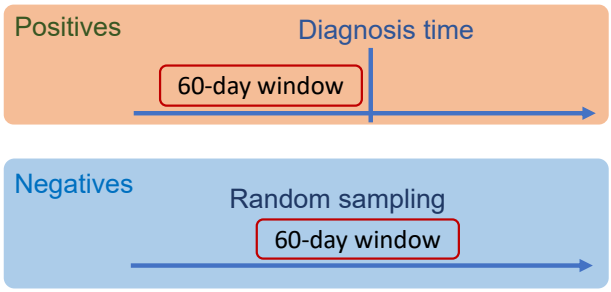


Figure 5.1: Wearable data sampling strategy (using a window size of 60 days as an example).

Table 5.1: Diagnosis distribution

Diagnosis Name	**Number of participants**
None	7749
Anxiety disorder	489
Major depression, single episode	295
Generalized anxiety disorder	134
Moderate recurrent major depression	59
Depressive disorder	44
Posttraumatic stress disorder	37
Dysthymia	36
Acute stress disorder	29
Mild recurrent major depression	23
Recurrent major depression	22
Panic disorder without agoraphobia	22
Mild major depression, single episode	15
Moderate major depression, single episode	12
Recurrent major depressive episodes, moderate	9
Recurrent depression	7
Chronic post-traumatic stress disorder	6
Mixed anxiety and depressive disorder	3
Panic disorder with agoraphobia	2
Recurrent major depressive episodes, mild	2
Recurrent major depressive episodes	1

in most populations were selected. This made our predictive model reach higher applicability. Second, characteristics that were strongly correlated with the diagnosis of depressive and anxiety disorders were excluded (e.g., mental health history), to avoid potential label leakages when building predictive models. As such, we included seven common characteristics: age, race, ethnicity, gender, education, alcohol history, and smoke history.

5.3.2 Statistical Analysis of Wearable Data

"All of Us" program provides the daily summary time-series data obtained from the Fitbit cloud, which includes 10 variables that are in a day-by-day granularity:

- **Total steps (totalSteps)**. Fitbit provides step measurements using the embedded inertial measurement unit (IMU). This is the total steps the user takes for one day.
- **Calories BMR (caloriesBMR)**. Total number of BMR calories burned for one day when the user is sedentary.
- **Calories out (caloriesOut)**. Total number of BMR calories burned for one day, associated with the activity, goal, summary totals. This variable includes the calories BMR and activity calories.
- **Fairly Active Minutes (fairlyActiveMinutes)**. Total minutes when the user is fairly/ moderately active for one day.
- **Marginal calories (marginalCalories)**. Total marginal estimated calories burned one the day.
- **Very active minutes (veryActiveMinutes)**. Total minutes the user is very active for one day.
- **Average heart rate (averageHeartRate)**. Average heart rate for one day.
- **Activity calories (activityCalories)**. introduced in Chapter 4.3.2.
- **Lightly active minutes (lightlyActiveMinutes)**. introduced in Chapter 4.3.2.
- **Sedentary minutes (sedentaryMinutes)**. introduced in Chapter 4.3.2.

To demonstrate possible differences in wearable data between the positive and the negative participants, we conducted statistical analysis on the daily summary variables. We adopted the

Table 5.2: Wearable Statistical features. Mean (S.D.) are reported per group.

Variables	Positive participants n = 1,247	Negative participants n= 7,749	Statistic	*p*-value
averageHeartRate_intercept	57.01 (32.74)	68.87 (17.72)	369.72	<0.005
caloriesOut_intercept	1718.64 (1094.34)	2173.42 (764.91)	342.24	<0.005
caloriesBMR_intercept	1110.12 (666.41)	1360.44 (405.82)	340.54	<0.005
sedentaryMinutes_intercept	590.21 (349.84)	717.30 (222.53)	299.79	<0.005
totalSteps_intercept	5099.27 (4210.90)	7245.23 (4214.18)	289.19	<0.005
averageHeartRate_std	11.80 (11.19)	7.65 (8.65)	234.52	<0.005
lightlyActiveMinutes_intercept	156.47 (116.03)	198.05 (85.79)	233.79	<0.005
activityCalories_intercept	717.51 (590.13)	961.77 (527.10)	230.97	<0.005
totalSteps_median	5038.14 (3914.17)	6874.21 (4109.60)	225.59	<0.005
caloriesOut_min	1198.72 (1023.89)	1614.52 (933.60)	214.85	<0.005
caloriesOut_median	1734.94 (982.75)	2088.54 (784.34)	209.33	<0.005
sedentaryMinutes_min	395.50 (315.52)	513.29 (272.17)	198.93	<0.005
marginalCalories_intercept	391.73 (359.36)	538.68 (346.78)	198.09	<0.005
caloriesBMR_std	215.14 (226.25)	137.80 (176.96)	195.06	<0.005
totalSteps_min	3123.46 (3481.09)	4726.41 (3896.07)	194.42	<0.005
caloriesBMR_min	803.24 (662.77)	1046.95 (572.98)	192.3	<0.005
averageHeartRate_min	40.37 (32.68)	51.89 (27.10)	188.88	<0.005
averageHeartRate_median	57.92 (28.42)	66.42 (19.09)	187.93	<0.005
sedentaryMinutes_median	591.48 (295.92)	686.77 (221.17)	185.73	<0.005
caloriesBMR_median	1131.57 (589.98)	1315.51 (430.81)	180.38	<0.005
activityCalories_min	443.55 (479.31)	639.78 (493.24)	177.92	<0.005
activityCalories_median	711.82 (550.64)	913.32 (513.72)	167.92	<0.005
marginalCalories_min	235.60 (279.10)	349.04 (299.13)	163.43	<0.005
lightlyActiveMinutes_min	99.90 (100.86)	134.71 (90.79)	158.53	<0.005
sedentaryMinutes_std	140.81 (122.04)	101.50 (101.63)	156.76	<0.005
lightlyActiveMinutes_median	156.02 (104.70)	188.61 (85.27)	151.74	<0.005
marginalCalories_median	387.19 (341.10)	509.92 (333.37)	150.08	<0.005
caloriesOut_std	358.27 (331.87)	259.39 (265.58)	142.94	<0.005
veryActiveMinutes_intercept	12.20 (20.76)	20.79 (25.64)	131.92	<0.005
veryActiveMinutes_median	11.61 (19.47)	19.26 (24.05)	118.94	<0.005
averageHeartRate_slope	0.04 (0.72)	-0.12 (0.45)	104.35	<0.005
caloriesBMR_slope	0.63 (13.81)	-2.31 (8.86)	101.53	<0.005
veryActiveMinutes_min	4.93 (13.09)	10.22 (18.33)	100.13	<0.005
caloriesOut_slope	0.76 (21.53)	-3.71 (14.62)	89.29	<0.005
fairlyActiveMinutes_intercept	11.85 (17.25)	17.24 (19.44)	88.23	<0.005
lightlyActiveMinutes_std	38.99 (29.66)	31.81 (25.55)	83.95	<0.005
sedentaryMinutes_slope	0.38 (7.90)	-1.18 (5.49)	77.21	<0.005
fairlyActiveMinutes_median	11.21 (16.43)	15.67 (17.54)	73.27	<0.005
fairlyActiveMinutes_min	4.90 (10.95)	8.02 (12.43)	72.78	<0.005
lightlyActiveMinutes_slope	0.05 (2.09)	-0.35 (1.60)	62.55	<0.005
activityCalories_slope	0.14 (10.09)	-1.71 (8.16)	53.25	<0.005
totalSteps_slope	0.83 (71.98)	-12.46 (63.64)	46.71	<0.005
marginalCalories_slope	0.07 (5.85)	-0.96 (4.97)	45.68	<0.005
veryActiveMinutes_std	5.06 (5.73)	6.25 (6.03)	44.57	<0.005
activityCalories_std	186.09 (149.12)	164.69 (132.14)	28.13	<0.005
fairlyActiveMinutes_slope	0.00 (0.29)	-0.03 (0.28)	16.16	<0.005
fairlyActiveMinutes_std	5.03 (5.11)	5.62 (5.41)	13.43	<0.005
veryActiveMinutes_slope	-0.00 (0.31)	-0.03 (0.33)	12.22	<0.005
marginalCalories_std	106.39 (90.31)	98.98 (82.92)	8.68	<0.005
totalSteps_std	1357.31 (1032.97)	1318.56 (979.65)	1.72	0.19

hierarchical feature engineering approach in Chapter 4.3.3 to transform each daily summary time series into five statistical features (i.e., median, min, max, slope, intercept). Then we used the one-way Analysis of variance (ANOVA) test on each statistical feature between the positive and the negative groups. Table 5.2 lists all the transformed wearable statistical features. Most of the extracted wearable features show significant differences between the positive and the negative (one-way ANOVA, $p-$value < 0.005), suggesting the potential discrimination power of wearable data when assessing the depressive and anxiety disorders at a group level.

5.3.3 Statistical Analysis of Static Characteristics

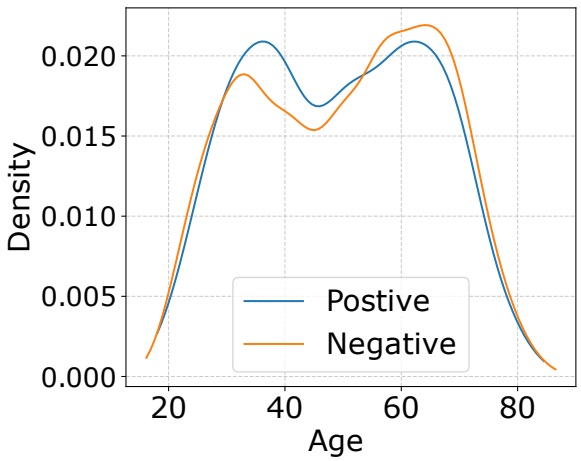


Figure 5.2: Age distribution (the positives and the negatives are normalized respectively).

Unlike previous studies [16, 30, 45, 113, 189, 209, 217] with a cohort from a single population or having less than 100 participants, our study has a wide spectrum of participants. Figure 5.2 displays the age distributions. Our study cohort has a wide range of ages from 15 to 90, with an average age of 48.6 (S.D. 15.9) for the entire cohort. We observed two peaks for both the positive and negative, which is at around age 35 and age 65. We hypothesize

that the younger (around age 35) are more fond of wearables due to fashion statement, and that the older (around age 65) adopt more wearables due to the growing need for health monitoring. A significant difference in mean age was found between the positive and the negative (Wilcoxon rank-sum test, $p = 0.002$).

An overview of all static characteristics is presented in Table 5.3. Among all the participants, 69.2% were female, 28.9% were male and 1.9% were not specified, demonstrating that females are more likely to take part in this study with wearables. Also, the female tends to be more vulnerable to depressive and anxiety disorders (χ^2 test of independence, $p < 0.005$). Previous literature [130] exhibited a similar observation. In terms of the race in the whole dataset, 83.4% were White, 5.0% were Black or African American, 3.2% were Asia, 5.4% were none of those (i.e., White, Black, Asia), 2.1% were more than one races and 0.7% skipped the question. Even though our study cohort is heavily skewed to the White, the absolute number of participants in each category is still considerably large compared to previous studies [16, 30, 45, 113, 189, 209, 217]. The mental disorders show a significant correlation to the race (χ^2 test of independence, $p < 0.005$). As for the ethnicity in the whole dataset, 91.9% were not Hispanic or Latino, 6.6% were Hispanic or Latino and 1.5% skipped the question. No significant correlation between the diagnosis and the ethnicity was found (χ^2 test of independence, $p = 0.790$). In the light of participants' highest education levels, 72.7% of the participants had college graduate or advanced degrees, 21.3% had college associate or technical school degrees, 5.0% had Grade 12 or high school degrees, 0.3% had below high school degrees and 0.3% had skipped the question. A significant difference in education level was observed between the positive and the negative (χ^2 test of independence, $p < 0.005$). On average, the negative group concentrates more on the college graduate or above degree when compared to the positive group. In addition to the basic demographic characteristics, we incorporated two other static characteristics: drink frequency and smoke

frequency. Comorbidity of alcohol use with some mental disorders is well established [170]. In the general population, the overlap between alcohol abuse/dependence and broadly classified depressive disorders is greater than expected by chance [170]. Also, smoke dependence demonstrates a similar link to depression [155]. In our analysis, we found both the drinking frequency and smoke frequency show significant differences between the positive and the negative (χ^2 test of independence, $p < 0.005$).

Table 5.3: Participant static characteristics

Static Characteristics	Positive participants n = 1,247	Negative participants n= 7,749	p-Value
Age, years, mean ± SD	47.3 ± 15.3	48.8 ± 15.9	<0.005
Gender, n (%)			<0.005
Female	1045 (80.1)	5185 (67.4)	
Male	232 (17.8)	2366 (30.8)	
Not specified	27 (2.1)	141 (1.8)	
Race, n (%)			<0.005
White	1127 (86.4)	6381 (83.0)	
Black or African American	53 (4.1)	398 (5.2)	
Asian	20 (1.5)	268 (3.5)	
None of these	69 (5.3)	421 (5.5)	
More than one population	25 (1.9)	167 (2.2)	
Skipped	10 (0.8)	57 (0.7)	
Ethnicity, n (%)			0.790
Not Hispanic or Latino	1205 (92.4)	7065 (91.8)	
Hispanic or Latino	81 (6.2)	514 (6.7)	
Skipped	18 (1.4)	113 (1.5)	
Education, n (%)			<0.005
College graduate or above	815 (62.5)	5734 (74.5)	
College One to Three	370 (28.4)	1555 (20.2)	
Twelve Or GED	106 (8.1)	348 (4.5)	
Below high school	9 (0.7)	26 (0.3)	
Skipped	4 (0.3)	29 (0.4)	
Drink, n (%)			<0.005
4 or More Per Week	153 (11.7)	1130 (14.7)	
2 to 3 Per Week	196 (15.0)	1395 (18.1)	
2 to 4 Per Month	322 (24.7)	1903 (24.7)	
Monthly Or Less	432 (33.1)	2095 (27.2)	
Never	163 (12.5)	778 (10.1)	
Skipped	38 (2.9)	391 (5.1)	
Smoke, n (%)			<0.005
Every Day	46 (3.5)	144 (1.9)	
Some Days	33 (2.5)	122 (1.6)	
Not At All	408 (31.3)	1983 (25.8)	
Skipped	817 (62.7)	5443 (70.8)	

5.4 Predictive Models

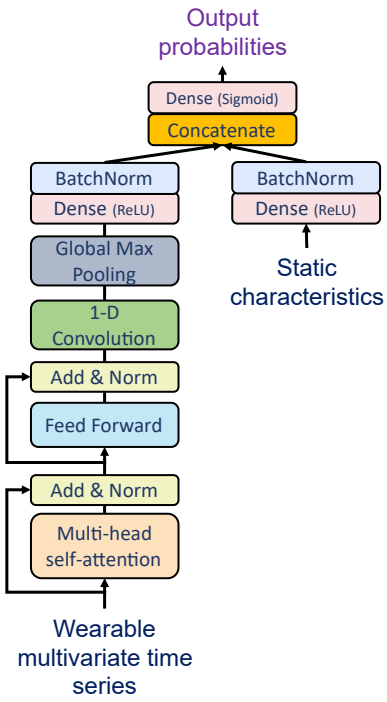


Figure 5.3: WearNet model architecture.

In this section, we discussed machine learning models to predict the defined depressive and anxiety disorders, using wearable data and static characteristics. We presented *WearNet*, a novel deep model combining a transformer encoder with a convolutional neural network, which can efficiently handle the long time-series sequences and show superior performances. WearNet comprises three key components: (1) a transformer encoder that identifies important patterns across multiple timestamps; (2) a convolutional layer that integrates neighborhood patterns; and (3) a global max-pooling layer that captures the overall patterns for the final probability prediction. Figure 5.3 illustrates the architecture of the WearNet.

Transformer Encoder

The inputs of the wearable data are in the form of multivariate time series, which usually have a length ranging from tens to hundreds of steps, depending on sampling frequency and window size. Recurrent neural network (RNN) has been previously used to handle time-series data as it can potentially capture positional and semantic information. But RNN suffers from the computational complexity and vanishing gradient problem, especially for long sequences[168]. Recently, the attention mechanism has become an alternative to the recurrent neural network, showing abilities to capture dependencies of various ranges (e.g., shorter-range vs. longer-range) within a sequence. In our proposed model, we utilized the transformer encoder [205], including the multi-head self-attention mechanism, to distill the raw wearable information.

Suppose we have a multivariate wearable time-series input $x_{wear} \in \mathbb{R}^{T \times D}$, the self-attention is calculated as following:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{KQ^T}{\sqrt{d_k}})V \tag{5.1}$$

$$Q = x_{wear}W^Q, K = x_{wear}W^K, V = x_{wear}W^V \tag{5.2}$$

Where Q, K, V represent the queries, keys, and values, respectively; W^Q, W^K and W^V are linear projection parameter matrices; d_k is the dimension of the keys. Then, we have a multi-head operation, which is basically performing the single self-attention multiple times. Multi-head self-attention allows the model to jointly attend to information from different representation subspaces at different positions.

$$h = \text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{5.3}$$

$$\text{where head}_\text{i} = \text{Attention}(xW_i^Q, xW_i^K, xW_i^V) \tag{5.4}$$

Where the W_i^Q, W_i^K and W_i^V are linear projection parameter matrices for the i-th head; h is the output of the multi-head self-attention operation. In our work, we employed 16 heads. After the multi-head self-attention, we added the position-wise feed-forward network (FFN) with a ReLU activation:

$$FFN(h) = \max(0, hW_1 + b_1)W_2 + b_2 \tag{5.5}$$

Where W_1, W_2, b_1 and b_2 are kernel weights and bias in the feed-forward network. There are shortcut connections (see Figure 5.3) for both the multi-head self-attention and feed-forward network to avoid being trapped by spurious local optimum while training [59].

Convolutional Neural Network

We found that adding a 1-dimensional convolutional layer promotes predictive performance (see Section 5.5.5). The convolutional layer contains a set of filters, which is to capture the patterns from neighborhood time steps. In our work, we have 10 filters with a kernel size of 3 and a step size of 1.

Global Maximum Pooling

The global maximum pooling enables the model to select one outstanding global pattern from all timestamps. It reduces the time-series feature maps into a single instance of feature map (without temporal dimension), by having the maximum feature value in each feature channel from all the temporal information. There is no parameter to optimize in the global maximum pooling layer, thus making it a robust transformation to avoid overfitting.

Integration of Wearable and Static Information

After the global maximum pooling layer, we further condense the feature map into an 8-dimensional vector with a dense layer and a batch normalization. Similarly, the static characteristics are also condensed into an 8-dimensional vector with a dense layer and a batch normalization. Then, we concatenated the two 8-dimensional vectors into a single 16-dimension vector, followed by a dense layer and a sigmoid activation to generate the probability estimates of a person having the defined depressive and anxiety disorder.

5.5 Experimental Evaluation

In this section, we evaluated the proposed WearNet from three perspectives: (1) comparison of predictive performance with other state-of-the-art models, (2) influence of missing data imputation, and (3) influence of the window size of wearable data. More importantly, we utilized a model-agnostic explanation tool to shed light on the prediction from the WearNet, which demonstrates the relative feature importance and the feature contribution directions (e.g., this feature makes the model tend to have a positive prediction). All the experiments were conducted on the dataset introduced in Section 5.3.

5.5.1 Evaluation Setting

We employed a time-based train/validation/test splitting scheme that is closest to practical clinical application scenarios. All the participants were ranked based on the first date in their wearable data window. The oldest 80% of participants were used during training and validating, out of which 10% were for validating. And the rest 20% of the participants were used during testing.

It is worth noting that the test dataset was imbalanced, with only 13.4% positives. Therefore, we used the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) as the major metrics. Those two metrics can well gauge the performance with the imbalanced dataset when the positive outcomes are of more interest. All the experiments were run 10 times, and the average performances with standard deviations were reported.

5.5.2 Comparing with Baselines

In this experiment, we evaluated the predictive performance of WearNet in comparison to a set of state-of-the-art models. We imputed the wearable data with an indicator variable (i.e., -1), except for those models designed for time-series prediction and imputation simultaneously. The window size is 60 days with a yield threshold of 10%.

The first category of the baseline is shallow machine learning models with traditional feature engineering techniques [106], including logistic regression (LR) with L2 regularization, random forest (RF), gradient boosting decision trees (GBDT), and support vector machine (SVM) with radial basis function kernel. Basically, we employed a one-dimensional concatenation of

the wearable statistical features listed in Table 5.2 and static characteristics listed in Table 5.3 as the input for those shallow machine learning models.

The second category of baselines is deep learning-based models, including the followings:

- **bidirectional long short-term memory (bi-LSTM)**:a basic bidirectional recurrent neural network (RNN) model with LSTM unit.
- **BRITS** [17]: a bidirectional RNN model with a built-in imputation component for handling missing values in the input. BRITS learns the missing values in a bidirectional recurrent dynamical system, without any specific assumption.
- **CrossNet** [105]: a bidirectional RNN model with a built-in imputation component that integrates static and time-series clinical data in deep recurrent models through multi-modal fusion.
- **Temporal convolutional network (TCN)** [133]: temporal convolutional network with dilation designed to handle long input sequence of time-series data.
- **Informer** [222]: a computationally efficient transformer-type model for long sequence predictions. We only utilized one encoder from Informer as we target at a classification task.
- **WearNet**: our proposed deep neural network with a combination of a transformer encoder and a convolutional neural network.

We employed the same concatenation components to integrate the static characteristics for the deep models, except CrossNet. CrossNet has its own mechanism to learn from the static characteristics via multi-modal fusion and static hidden state initialization [105]. All the

models were implemented in TensorFlow [118] or Scikit-learn [145] Python framework. The hyperparameters were tuned with grid search.

Table 5.4: Predictive performances of all models

Category	Model	AUROC	AUPRC
Shallow Models with feature engineering	LR	0.701(0.000)	0.351(0.000)
	SVM	0.592(0.000)	0.290(0.000)
	RF	0.661(0.005)	0.349(0.007)
	GBDT	0.685(0.001)	0.365(0.000)
Deep Models	bi-LSTM	0.702(0.015)	0.464(0.011)
	BRITS	0.693(0.012)	0.445(0.011)
	CrossNet	0.682(0.021)	0.429(0.014)
	TCN	0.629(0.021)	0.235(0.024)
	Informer	0.705(0.008)	0.428(0.011)
	WearNet	0.717(0.009)	0.487(0.008)

Table 5.4 shows the performances of all the models. It is obvious that WearNet shows the best performance, suggesting the effectiveness of our proposed architecture to learn from the wearable time-series data and the static characteristics. bi-LSTM demonstrates better performances than BRITS and CrossNet, even though they are all RNN-based models. CrossNet shows worse performance, possibly due to the fact that CrossNet focuses on the multi-modal fusion to boost performance which needs more information from the static features, but our dataset has limited statistic features. TCN shows the worst performance among the deep models, indicating the temporal convolution architecture is not suitable to distill information from the wearable time-series data in our dataset. Informer demonstrates a similar AUROC but a lower AUPRC compared to WearNet. Both Informer and WearNet are attention-based models, but Informer utilized a sampling strategy to lower the computational complexity of the attention mechanism. When compared to the shallow models with feature engineering, all the deep learning models show superior performances, indicating the advantages of deep learning

on the large dataset. The feature engineering approach requires significant efforts on the hand-crafted features, and is hard to comprehensively capture the underlying inter-individual differences.

5.5.3 Impacts of Imputation Values

The missing data is inevitable for the wearables. There are various factors that can lead to missing data. For example, the device needs to charge or the participants forget to wear the device. Most machine learning models need intact inputs without missing values. Both BRITS and CrossNet have dedicated components to handle the missing values in the time-series data, using data-driven imputation approaches. BRITS utilized the correlations from history values and other time-series variables for the imputation, while CrossNet additionally incorporated the static characteristics (e.g., demographic characteristics and other clinical lab tests) for the imputation. Nonetheless, those dedicated components did not promote the performance compared to the WearNet with a fixed value (i.e., -1) imputation (see Table 5.4).

Moreover, we evaluated another common imputation strategy that is to use personal mean value to fill the missing value, in comparison to the fixed value (i.e., -1) imputation used in the previous experiment. Table 5.5 shows the performance comparisons. The mean value imputation did not show improved performances either.

Table 5.5: Imputation impacts

Model	AUROC	AUPRC
WearNet(mean impute)	0.664(0.006)	0.231(0.008)
WearNet(-1 impute)	0.717(0.009)	0.487(0.008)

Both the mean value imputation and data-driven imputation (e.g., BRITS and CrossNet) might ignore the missing patterns. Those patterns could be suggestive of mental disorders. For example, the participants were staying at home and not wearing the activity tracker, thus incurring missing values. The staying-at-home event itself may be a predictor. If we use some values from the imputation methods to fill in the missing values, spurious information will be introduced and the model may treat these values as normal, ignoring the missing patterns. As such, it is reasonable to just add an indicator (e.g., -1) to represent the missingness. The above results also proved that our proposed WearNet can effectively handle the missing data via a simple fixed indicator, which shows even better performances than the deep model with dedicated imputation components.

5.5.4 Impacts of Window Size

Previously, we evaluated the models with a fixed window size of 60 days, given the long-term characteristics of the mental disorders. To validate the feasibility of the proposed model with different window sizes, we conducted experiments with varying window sizes. The yield threshold is still fixed at 10%. Figure 5.4 illustrates the performances with varying window sizes.

We can observe that the performances increase when the window size varies from 15 days to 90 days, which is especially obvious from 15 days to 60 days. This trend indicates that a relatively large window size (e.g., from 60 days to 90 days) indeed helps in predicting the defined depressive and anxiety disorders. However, when the window size becomes even larger (from 90 days to 120 days), the performances drop a bit. We hypothesize that it is due to the fact that the model may not well process the long sequences. A larger window size not only incurs more model complexity, but also makes the prediction less "real-time", as it requires a longer period to collect the data. As such, it is reasonable to choose 60 days

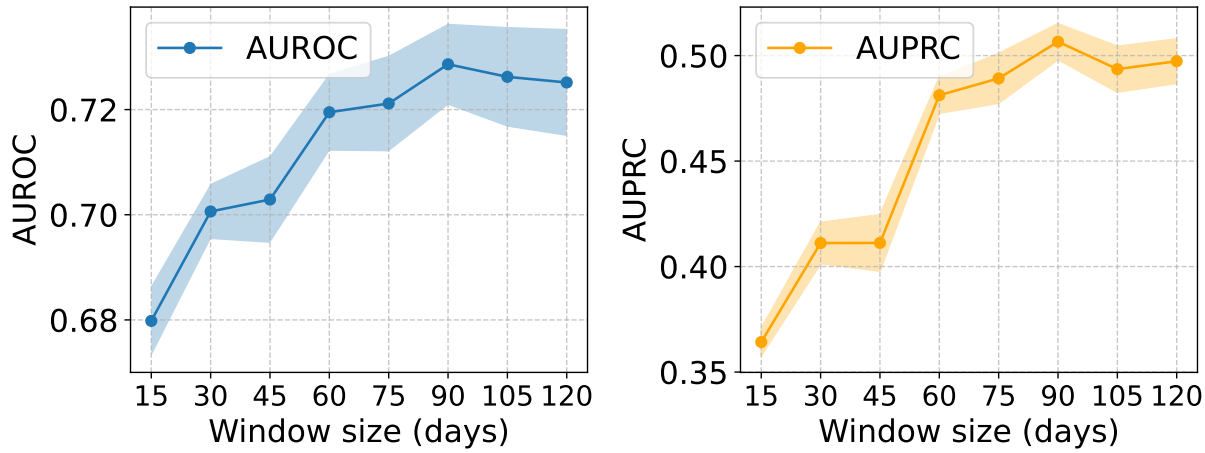


Figure 5.4: WearNet performances with varying window sizes

as the window size in our primary analysis, given the trade-off between the performance and many constraints of the large window size.

5.5.5 Ablation Study

Our WearNet has an additional convolutional layer compared to a traditional transformer encoder. To support the performance gain from the convolutional layer, we evaluated our model without the convolutional layer. Besides, we also evaluated our model with different sensing modalities. Table 5.6 shows the performances of WearNet with different settings. We can observe that adding a convolutional layer after the transformer encoder indeed boosted the performance. Dropping the static characteristics makes the WearNet have a slight performance decline, and performance declines even more when dropping the wearable data. This observation attests that the wearable data played a more significant role than the static data in our model, but combining both achieved the best performances.

Table 5.6: Ablation study performances

Model	AUROC	AUPRC
WearNet	0.717(0.009)	0.487(0.008)
WearNet(without convolutional layer)	0.673(0.008)	0.251(0.008)
WearNet(without static characteristics)	0.702(0.007)	0.456(0.007)
WearNet(without wearable data)*	0.650(0.003)	0.222(0.003)

*The model reduces to a single-layer perceptron if we drop components for wearable data.

5.5.6 Model Explanation

We utilized the integrated gradients (IGs) [198], an explainable AI technique that aims to explain the relationship between a model's prediction in terms of its input values. The IG assigns an importance score to each input value by approximating the integral of gradients of the model's output along the linear path from given references to the inputs [198]. We randomly sampled reference values from the training dataset, and we calculated the expectations of IGs for each of the inputs in our testing dataset.

As the wearable data are in the form of multivariate time series, we were interested in the aggregated IGs for the whole time series instead of the single value in each timestamp. Those aggregated IGs will be used as relative feature importances. For example, we aggregated IGs for daily step time series by averaging the calculated absolute gradients in each time step, then compared the aggregated IGs with other time series (e.g., activity calories). Figure 5.5 demonstrates the ranking of feature importance quantified by the aggregated IGs. The total step (i.e., "totalSteps") time series is the most important in our model for detecting the defined depressive and anxiety disorders. Previous literature has confirmed the predictive power of the daily steps [32, 121, 178]. Other time series show moderate importance, including calories out, calories BMR, activity calories and sedentary minutes. The very active minutes time series shows the least importance, which could possibly owe to the fact that the participants

only have an occasional period that was classified as very active by the Fitbit [214], thus making it too sparse to contain useful information for the model.

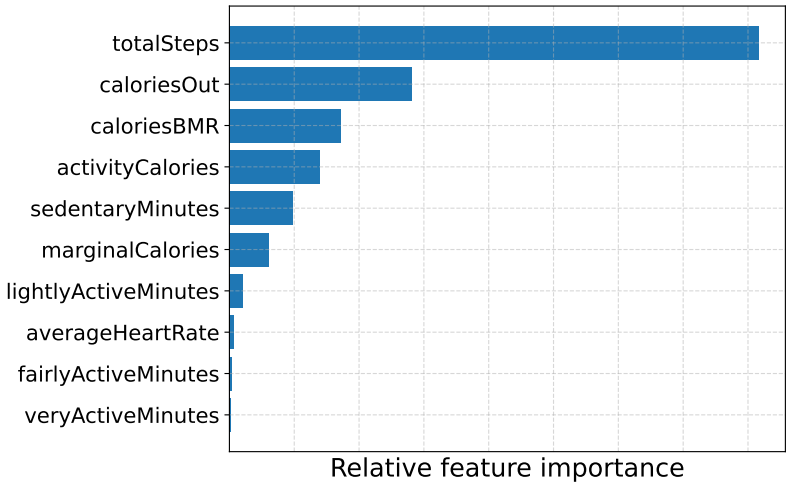
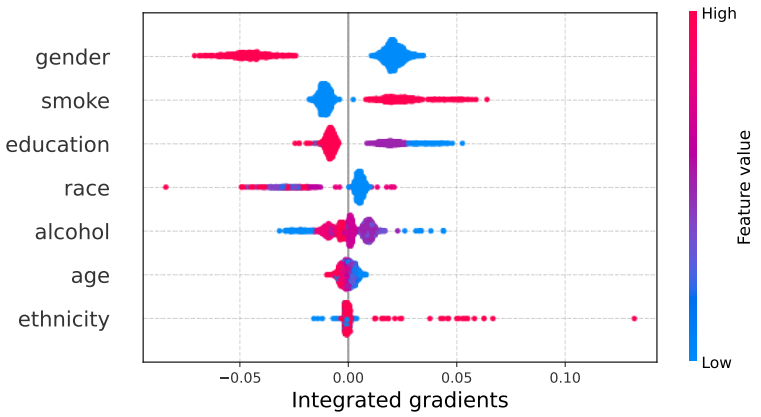


Figure 5.5: Wearable time series feature importance

For the static characteristics, we directly adopted the IGs for each value as they are one-dimensional without the temporal dimension. The positive IG means that the value increases the probability of the positive prediction (i.e., having the defined mental disorders), whereas the negative IG means that the value decreases the probability. Figure 5.6 demonstrates the IGs for the static characteristics. Different colors signify different values of that characteristic. For example, the blue in gender signifies the woman, and the red signifies the man. If the characteristic can be ranked, the red color represents a higher value. For example, the red in smoke represents a higher smoke frequency. It is obvious that the model tends to predict a woman having a higher probability of mental disorders, which has been proved in previous literature [130]. People who smoke more frequently also tend to have a higher probability of mental disorders. However, for the race, alcohol, age and ethnicity, there seem to be no clear separations or unique patterns based on the IGs.



*__Gender__ from blue to red: female, male and not specified; __Smoke__ from blue to red: skipped, not at all, some days, every day; __Education__ from blue to red: below high school, twelve Or GED, College One to Three, College graduate or above, skipped; __Race__ from blue to red: White, Asian, Black or African American, more than one population, none of those, skipped; __Alcohol__ from blue to red: skipped, never, monthly or less, 2 to 4 per month, 2 to 3 per week, 4 or more per week; __Age__ from blue to red: 15 to 85; __Ethnicity__ from blue to red: Hispanic or Latino, not Hispanic or Latino, non of these, skipped

Figure 5.6: Static characteristic feature importance

5.6 Discussion and Conclusion

Depressive and anxiety disorders are usually long-term disorders, which could span for several months or even several years. Although highly prevalent, those mental disorders are usually hard to identify due to numerous factors. Current gold-standard assessment relies on clinical visits by means of questionnaires, which usually demands a large amount of time and money. Passive monitoring with wearables could greatly help alleviate those burdens.

In this work, we explored detecting depressive and anxiety disorders with wearables on a large public dataset. We proposed a novel deep neural network that combines the transformer encoder and convolutional neural network, which can effectively distill information from time-series data. We evaluated our model on a large cohort consisting of more than 11,000 participants. The model primarily takes 60-day wearable data and basic static characteristics as inputs, achieving an AUROC of 0.717 (S.D. 0.009) and an AUPRC of 0.487 (S.D. 0.008), which outperforms other baseline models. We also evaluated the impacts of imputation

techniques, demonstrating the effectiveness of the fixed indicator imputation. Fixed indicator imputation can encode the information of the patterns when there are missing values. The missing pattern itself may reflect the information of the user habits. We found that the input window size of the wearable data has an impact on the performances as well. A larger window size would improve the model, but the gain drops when the window size increases beyond 90 days. There is a trade-off between the model performance and the model constraints. Therefore, we chose the 60-day window. More importantly, we quantified the contributions of the data modalities using the ablation study and model explanation tool. In terms of the model performances, the wearable data shows more predictive power than the static characteristics, but combining them achieves the best. The total step time series account for the most importance among the wearable data. Gender and smoke history are the top two important static characteristics, which suggests that females and frequent smokers tend to be more vulnerable to depressive and anxiety disorders. Those findings in the model explanation match our statistic analysis, verifying the rationality of the predictions from our proposed model.

To the best of our knowledge, our study is based on the largest and most diverse cohort, and demonstrates decent predictive performances. Wide accessible, and not intrusive or burdensome, our approach with wearables represents a promising step in discovering depressive and anxiety disorders in public, in complementary to the traditional diagnostic tools from healthcare providers.

5.7 Limitation

We note several limitations in this work. First, even though our dataset is the largest dataset with the wearables for studying depressive and anxiety disorders, there is an inherent bias in

the dataset from the "All of Us" program. In Section 5.3, we observed the cohort has more females than males, and the race, as well as ethnicity, are also not evenly distributed. Second, we did not include the minute-by-minute heart rate and step time series in our analysis. Those fine-grained time series could boost the model performances if with an effective approach. Third, the dataset only contains the wearable data from Fitbit devices. The applicability of other wearables remains to be investigated.

Chapter 6

Conclusion

The advancements in wearables pose new research challenges in the area of smart sensing and clinical outcome predictions. This dissertation has studied the applications of wearables in various scenarios from physiological signal measurements to mental health predictions. Those applications constructed a promising path toward the practical adoption of wearables in healthcare and clinical monitoring, solving challenges at different levels.

RespWatch (in Chapter 2) investigated the low-level respiratory rate measurements from a commercial smartwatch, using a hybrid approach combining deep learning and signal processing. The signal processing shows high accuracy under moderate noise, whereas deep learning shows more robustness to significant noise. Our hybrid approach dynamically leverages the complementary advantages for both signal processing and deep learning. However, the efficiency of the hybrid approach is still not comparable to the signal processing techniques. And our user study was limited to measuring respiratory rate under some types of activities. Further optimizations of computational resources and evaluations of applicability under more

circumstances are needed to unleash the full potential of measuring respiratory rate and monitoring respiratory-related disease with a smartwatch.

In Chapter 3, we studied stress detection with the RespWatch pipeline, by comparing the performances of the machine learning models in detecting subjective and objective stress. Stress is a much more complicated physiological and/or psychological response compared to respiratory rate. Long-term stress is even labeled as a "silent killer". The systematical evaluations of stress detection models in our study provide some preliminary understandings of the interplay between subjective and objective markers of stress. Nonetheless, we acknowledge several limitations of this study. This was a single-site study with 32 participants, and as such the results may not be generalizable. Future larger studies with more people are needed, especially in a free-living setting.

Following the detection of stress, Chapter 4 explored using wearables and multi-task learning (MTL) techniques in predicting the depression outcome along with integrated behavior therapy in a randomized controlled trial (RCT). Insights from RCTs are often translated into clinical practice. Our idea helps in streamlining the clinical point-of-care use of an already successful intervention by considering patient-specific characteristics such as baseline clinical, and wearable-device-based activity characteristics. The application of MTL techniques to RCTs is new and provides a new frontier for precision treatment. While this approach is powerful and useful in assessing the value of an intervention for clinical practice, our model does not directly help in assessing "which patient" should receive an intervention. Our model relied on the baseline clinical characteristics and 2-month's wearable data, which assumed that the patient's treatment plan (i.e., intervention or control) was already known. For behavioral interventions, sometimes outcomes are achieved without any treatment (e.g., a wait-and-watch approach). For example, some participants in the control group would show some depression outcome improvement without the treatment intervention. In the

future, we may build the MTL model, purely utilizing the baseline data that does not depend on any information on treatment assignments, so that we can help in deciding whether a patient should receive the treatment at the point of care. This not only helps in intervention choice decisions, but also in potentially changing the frequency/dose (e.g., number of times a particular therapy must be used) of an intervention.

Finally, in Chapter 5, we investigated the applicability of wearables in detecting depressive and anxiety disorders with a large cohort consisting of a wide spectrum of populations. Although highly prevalent, those mental disorders are usually hard to identify due to numerous factors. We can passively employ our detection model to assist in the diagnoses of mental disorders in the general public. The findings in this study could potentially benefit more people compared to previous studies, as the study cohort has fewer restrictions. The wearable activity brings minimum burdens to the users, as it has already been a part of daily life for many people. However, cautions for bias are still needed when applying the model and analyzing the results. First, our training data is skewed to female and White people. Even though we have a considerable number of people in other gender or races, the model may have a potential incline to the majority. Second, the labels in the dataset are far from perfect. Some people may not get diagnosed, and the diagnosis time from the electronic health records may not be truly accurate, as some patients may get a delayed diagnosis.

6.1 Closing Remarks

While the application studies in this thesis afford a promising path toward the practical adoption of wearables in healthcare, there are still some limitations we need to address, such as the inherent bias of the training dataset, small sample size, and applicability under more complex circumstances. Within the foreseeable future, the development of hardware and

software platforms, and wearable-based artificial intelligence analytic tools will help us to address those limitations and boost more applications of wearables in healthcare. We believe the wearables will continue to meet the emerging demands, and we will observe a more significant role of wearables in precision medicine with personalized healthcare solutions.

References

[1] *1.11. Ensemble methods — scikit-learn 1.0.2 documentation*. (Accessed on 04/17/2020).

[2] *All of Us Research Program Expands Data Collection Efforts with Fitbit | National Institutes of Health (NIH) — All of Us*. (Accessed on 04/24/2022).

[3] Daniel Almirall, Inbal Nahum-Shani, Nancy E Sherwood, and Susan A Murphy. "Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research." In: *Translational behavioral medicine* 4.3 (2014), pp. 260–274.

[4] *Inspect energy use with Energy Profiler*. 2020. URL: `https://developer.android.com/studio/profile/energy-profiler`.

[5] Michael A Andrykowski, Matthew J Cordova, Jamie L Studts, and Thomas W Miller. "Posttraumatic stress disorder after treatment for breast cancer: Prevalence of diagnosis and use of the PTSD Checklist—Civilian Version (PCL—C) as a screening instrument." In: *Journal of consulting and clinical psychology* 66.3 (1998), p. 586.

[6] Sonal Arora, Nick Sevdalis, Debra Nestel, Maria Woloshynowych, Ara Darzi, and Roger Kneebone. "The impact of stress on surgical performance: a systematic review of the literature." In: *Surgery* 147.3 (2010), pp. 318–330.

[7] American Psychological Association et al. *Stress in America(2019)*. 2019.

[8] Sadam Al-Azani and El-Sayed M El-Alfy. "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text." In: *Procedia Computer Science* 109 (2017), pp. 359–366.

[9] Sarah Louise Bell, Suzanne Audrey, David Gunnell, Ashley Cooper, and Rona Campbell. "The relationship between physical activity, mental wellbeing and symptoms of mental health disorder in adolescents: a cohort study." In: *International Journal of Behavioral Nutrition and Physical Activity* 16.1 (2019), pp. 1–12.

[10] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health." In: *Psychiatric rehabilitation journal* 38.3 (2015), p. 218.

[11] Sourav Bhattacharya and Nicholas D Lane. "From smart to deep: Robust activity recognition on smartwatches using deep learning." In: *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE. 2016, pp. 1–6.

[12] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. "Pervasive stress recognition for sustainable living." In: *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE. 2014, pp. 345–350.

[13] Daniel J Buysse, Charles F Reynolds, Timothy H Monk, Susan R Berman, David J Kupfer, et al. "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research." In: *Psychiatry res* 28.2 (1989), pp. 193–213.

[14] Danilo Bzdok and Andreas Meyer-Lindenberg. "Machine learning for precision psychiatry: opportunities and challenges." In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3.3 (2018), pp. 223–230.

[15] Ashley E Cain, Colin A Depp, and Dilip V Jeste. "Ecological momentary assessment in aging research: a critical review." In: *Journal of psychiatric research* 43.11 (2009), pp. 987–996.

[16] Luca Canzian and Mirco Musolesi. "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis." In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 2015, pp. 1293–1304.

[17] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. "Brits: Bidirectional recurrent imputation for time series." In: *Advances in neural information processing systems* 31 (2018).

[18] Daniel G Carey. "Quantifying differences in the "fat burning" zone and the aerobic zone: implications for training." In: *The Journal of Strength & Conditioning Research* 23.7 (2009), pp. 2090–2095.

[19] Charles S Carver, Michael F Scheier, and Jagdish K Weintraub. "Assessing coping strategies: a theoretically based approach." In: *Journal of personality and social psychology* 56.2 (1989), p. 267.

[20] Marta E Cecchinato, Anna L Cox, and Jon Bird. "Smartwatches: the Good, the Bad and the Ugly?" In: *Proceedings of the 33rd Annual ACM Conference extended abstracts on human factors in computing systems*. 2015, pp. 2133–2138.

[21] Peter H Charlton, Timothy Bonnici, Lionel Tarassenko, David A Clifton, Richard Beale, and Peter J Watkinson. "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram." In: *Physiological measurement* 37.4 (2016), p. 610.

[22] Peter H Charlton et al. "Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review." In: *IEEE reviews in biomedical engineering* 11 (2017), pp. 2–20.

[23] Adam Mourad Chekroud et al. "Cross-trial prediction of treatment outcome in depression: a machine learning approach." In: *The Lancet Psychiatry* 3.3 (2016), pp. 243–250.

[24] Jenny Chum et al. "Acceptability of the Fitbit in behavioural activation therapy for depression: a qualitative study." In: *Evidence-based mental health* 20.4 (2017), pp. 128–133.

[25] Paul Ciechanowski, Naomi Chaytor, John Miller, Robert Fraser, Joan Russo, Jurgen Unutzer, and Frank Gilliam. "PEARLS depression treatment for individuals with epilepsy: a randomized controlled trial." In: *Epilepsy & Behavior* 19.3 (2010), pp. 225–231.

[26] Paul Ciechanowski et al. "Community-integrated home-based depression treatment in older adults: a randomized controlled trial." In: *Jama* 291.13 (2004), pp. 1569–1577.

[27] Sheldon Cohen. "Perceived stress in a probability sample of the United States." In: (1988).

[28] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. "A global measure of perceived stress." In: *Journal of health and social behavior* (1983), pp. 385–396.

[29] William Coryell, Jess G Fiedorowicz, David Solomon, Andrew C Leon, John P Rice, and Martin B Keller. "Effects of anxiety on the long-term course of depressive disorders." In: *The British Journal of Psychiatry* 200.3 (2012), pp. 210–215.

[30] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. "A review of depression and suicide risk assessment using speech analysis." In: *Speech Communication* 71 (2015), pp. 10–49.

[31] Thomas J D'Zurilla, Arthur M Nezu, and Albert Maydeu-Olivares. "Social problem-solving inventory-revised." In: (2002).

[32] F Dimeo, M Bauer, I Varahram, G Proest, and U Halter. "Benefits from aerobic exercise in patients with major depression: a pilot study." In: *British journal of sports medicine* 35.2 (2001), pp. 114–117.

[33] Ioannis Dologlou and George Carayannis. "Pitch detection based on zero-phase filtering." In: *Speech Communication* 8.4 (1989), pp. 309–318.

[34] Laurita Dos Santos, Joaquim J Barroso, Elbert EN Macau, and Moacir F de Godoy. "Application of an automatic adaptive filter for heart rate variability analysis." In: *Medical engineering & physics* 35.12 (2013), pp. 1778–1785.

[35] Jessilyn Dunn, Ryan Runge, and Michael Snyder. "Wearables and the medical revolution." In: *Personalized medicine* 15.5 (2018), pp. 429–448.

[36] Victoria B Egizio, Michael Eddy, Matthew Robinson, and J Richard Jennings. "Efficient and cost-effective estimation of the influence of respiratory variables on respiratory sinus arrhythmia." In: *Psychophysiology* 48.4 (2011), pp. 488–494.

[37] Mohamed Elsadek, Minkai Sun, Ryo Sugiyama, and Eijiro Fujii. "Cross-cultural comparison of physiological and psychological responses to different garden styles." In: *Urban forestry & urban greening* 38 (2019), pp. 74–83.

[38] Takeshi Emura, Shigeyuki Matsui, and Hsuan-Yu Chen. "compound. Cox: univariate feature selection and compound covariate for predicting survival." In: *Computer methods and programs in biomedicine* 168 (2019), pp. 21–37.

[39] Sherrill Evans, Sube Banerjee, Morven Leese, and Peter Huxley. "The impact of mental illness on quality of life: A comparison of severe mental illness, common mental disorder and healthy population samples." In: *Quality of life research* 16.1 (2007), pp. 17–29.

[40] Michael W Eysenck and Małgorzata Fajkowska. *Anxiety and depression: toward overlapping and distinctive features*. 2018.

[41] Michael P Fay and Michael A Proschan. "Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules." In: *Statistics surveys* 4 (2010), p. 1.

[42] Julio Fernandez-Mendoza, Sarah Shea, Alexandros N Vgontzas, Susan L Calhoun, Duanping Liao, and Edward O Bixler. "Insomnia and incident depression: role of objective sleep duration and natural history." In: *Journal of sleep research* 24.4 (2015), pp. 390–398.

[43] John F Fieselmann, Michael S Hendryx, Charles M Helms, and Douglas S Wakefield. "Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients." In: *Journal of general internal medicine* 8.7 (1993), pp. 354–360.

[44] Foroohar Foroozan, Madhan Mohan, and Jian Shu Wu. "Robust beat-to-beat detection algorithm for pulse rate variability analysis from wrist photoplethysmography signals." In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 2136–2140.

[45] Mads Frost, Gabriela Marcu, Rene Hansen, Karoly Szaántó, and Jakob E Bardram. "The MONARCA self-assessment system: Persuasive personal monitoring for bipolar patients." In: *2011 5th international conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops*. IEEE. 2011, pp. 204–205.

[46] Nicole B Gabler, Naihua Duan, Sunita Vohra, and Richard L Kravitz. "N-of-1 trials in the medical literature: a systematic review." In: *Medical care* (2011), pp. 761–768.

[47] Richard M Glass, Andrew T Allan, EH Uhlenhuth, Chase P Kimball, and Dennis I Borinstein. "Psychiatric screening in a medical clinic: An evaluation of a self-report inventory." In: *Archives of General Psychiatry* 35.10 (1978), pp. 1189–1195.

[48] Sarah M Goodday and Stephen Friend. "Unlocking stress and forecasting its consequences with digital technology." In: *NPJ digital medicine* 2.1 (2019), pp. 1–5.

[49] Sarah M Goodday and Stephen Friend. "Unlocking stress and forecasting its consequences with digital technology." In: *NPJ Digital Medicine* 2.1 (2019), pp. 1–5.

[50] *Wear OS by Google Smartwatches*. 2020. URL: `https://wearos.google.com/#hands-free-help`.

[51] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. "Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis." In: *Journal of medical Internet research* 21.11 (2019), e16273.

[52] Mark Andrew Hall et al. "Correlation-based feature selection for machine learning." In: (1999).

[53] Nils Yannick Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Plötz. "PD disease state assessment in naturalistic environments using deep learning." In: *Twenty-Ninth AAAI conference on artificial intelligence*. 2015.

[54] Tian Hao, Chongguang Bi, Guoliang Xing, Roxane Chan, and Linlin Tu. "MindfulWatch: A smartwatch-based system for real-time respiration monitoring during meditation." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–19.

[55] Eduardo Hariton and Joseph J Locascio. "Randomised controlled trials—the gold standard for effectiveness research." In: *BJOG: an international journal of obstetrics and gynaecology* 125.13 (2018), p. 1716.

[56] Juliet Hassard, Kevin RH Teoh, Gintare Visockaite, Philip Dewe, and Tom Cox. "The cost of work-related stress to society: A systematic review." In: *Journal of occupational health psychology* 23.1 (2018), p. 1.

[57] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures." In: *PloS one* 6.12 (2011), e28210.

[58] Douglas M Hawkins. "The problem of overfitting." In: *Journal of chemical information and computer sciences* 44.1 (2004), pp. 1–12.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[61] The Lancet Global Health. "Mental health matters." In: *The Lancet. Global Health* 8.11 (2020), e1352.

[62] Emily T Hébert et al. "A mobile Just-in-Time adaptive intervention for smoking cessation: pilot randomized controlled trial." In: *Journal of medical Internet research* 22.3 (2020).

[63] CJK Henry. "Basal metabolic rate studies in humans: measurement and development of new equations." In: *Public health nutrition* 8.7a (2005), pp. 1133–1152.

[64] Javier Hernandez, Daniel McDuff, and Rosalind W Picard. "Biowatch: estimation of heart and breathing rates from wrist motions." In: *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE. 2015, pp. 169–176.

[65] Alberto Hernando et al. "Inclusion of respiratory frequency information in heart rate variability analysis for stress assessment." In: *IEEE journal of biomedical and health informatics* 20.4 (2016), pp. 1016–1025.

[66] Alberto Hernando et al. "Inclusion of respiratory frequency information in heart rate variability analysis for stress assessment." In: *IEEE journal of biomedical and health informatics* 20.4 (2016), pp. 1016–1025.

[67] Blake Anthony Hickey et al. "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review." In: *Sensors* 21.10 (2021), p. 3461.

[68] Blake Anthony Hickey et al. "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review." In: *Sensors* 21.10 (2021), p. 3461.

[69] Melissa D Hladek, Sarah L Szanton, Young-Eun Cho, Chen Lai, Caroline Sacko, Laken Roberts, and Jessica Gill. "Using sweat to measure cytokines in older adults compared to younger adults: A pilot study." In: *Journal of immunological methods* 454 (2018), pp. 1–5.

[70] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. "cStress: towards a gold standard for continuous stress assessment in the mobile environment." In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 2015, pp. 493–504.

[71] Xin Hu, Rahav Dor, Steven Bosch, Anita Khoong, Jing Li, Susan Stark, and Chenyang Lu. "Challenges in studying falls of community-dwelling older adults in the real world." In: *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE. 2017, pp. 1–7.

[72] Hugh Hunkin, Daniel L King, and Ian T Zajac. "Perceived acceptability of wearable devices for the treatment of mental health problems." In: *Journal of clinical psychology* 76.6 (2020), pp. 987–1003.

[73] Sandra K Hunter. "Sex differences in human fatigability: mechanisms and insight to physiological responses." In: *Acta physiologica* 210.4 (2014), pp. 768–789.

[74] Sinh Huynh, Rajesh Krishna Balan, JeongGil Ko, and Youngki Lee. "VitaMon: measuring heart rate variability using smartphone front camera." In: *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 2019, pp. 1–14.

[75] Frank R Ihmig, Frank Neurohr-Parakenings, Sarah K Schäfer, Johanna Lass-Hennemann, and Tanja Michael. "On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals." In: *Plos one* 15.6 (2020), e0231517.

[76] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. "μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch." In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2016, pp. 1124–1128.

[77] Dan V Iosifescu, Scott Greenwald, Philip Devlin, David Mischoulon, John W Denninger, Jonathan E Alpert, and Maurizio Fava. "Frontal EEG predictors of treatment outcome in major depressive disorder." In: *European Neuropsychopharmacology* 19.11 (2009), pp. 772–777.

[78] Ali Jalali, Pradeep Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. "On learning discrete graphical models using group-sparse regularization." In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 378–387.

[79] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. "Multi-task, multi-kernel learning for estimating individual wellbeing." In: *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*. Vol. 898. 2015, p. 3.

[80] Delaram Jarchi, Dario Salvi, Lionel Tarassenko, and David A Clifton. "Validation of instantaneous respiratory rate using reflectance PPG from different body positions." In: *Sensors* 18.11 (2018), p. 3705.

[81] Houtan Jebelli, Byungjoo Choi, Hyeonseung Kim, and SangHyun Lee. "Feasibility study of a wristband-type wearable sensor to understand construction workers' physical and mental status." In: *Construction Research Congress*. 2018, pp. 367–377.

[82] Maurice Jetté, Ken Sidney, and G Blümchen. "Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity." In: *Clinical cardiology* 13.8 (1990), pp. 555–565.

[83] Zhenhua Jia, Amelie Bonde, Sugang Li, Chenren Xu, Jingxian Wang, Yanyong Zhang, Richard E Howard, and Pei Zhang. "Monitoring a person's heart rate and respiratory rate on a shared bed using geophones." In: *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 2017, pp. 1–14.

[84] Walter Karlen, J Mark Ansermino, and Guy Dumont. "Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications." In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2012, pp. 3131–3134.

[85] Walter Karlen, Srinivas Raman, J Mark Ansermino, and Guy A Dumont. "Multiparameter respiratory rate estimation from the photoplethysmogram." In: *IEEE Transactions on Biomedical Engineering* 60.7 (2013), pp. 1946–1953.

[86] Chandan Karmakar, Ahsan Khandoker, Thomas Penzel, Christoph Schöbel, and Marimuthu Palaniswami. "Detection of respiratory arousals using photoplethysmography (PPG) signal in sleep apnea patients." In: *IEEE journal of biomedical and health informatics* 18.3 (2013), pp. 1065–1073.

[87] Palanisamy Karthikeyan, Murugappan Murugappan, and Sazali Yaacob. "A review on stress inducement stimuli for assessing human stress using physiological signals." In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. IEEE. 2011, pp. 420–425.

[88] Martin B Keller. "Depression: a long-term illness." In: *The British Journal of Psychiatry* 165.S26 (1994), pp. 9–15.

[89] Alex Kendall, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7482–7491.

[90] Ronald C Kessler and Alex Luedtke. "Pragmatic Precision Psychiatry—A New Direction for Optimizing Treatment Selection." In: *JAMA psychiatry* (2021).

[91] Jinhyuk Kim, Toru Nakamura, Hiroe Kikuchi, Tsukasa Sasaki, and Yoshiharu Yamamoto. "Co-variation of depressive mood and locomotor dynamics evaluated by ecological momentary assessment in healthy humans." In: *PLoS One* 8.9 (2013), e74979.

[92] Jungyoon Kim, Jangwoon Park, and Jaehyun Park. "Development of a statistical model to classify driving stress levels using galvanic skin responses." In: *Human Factors and Ergonomics in Manufacturing & Service Industries* 30.5 (2020), pp. 321–328.

[93] Meelim Kim, Jaeyeong Yang, Woo-Young Ahn, Hyung Jin Choi, et al. "Machine Learning Analysis to Identify Digital Behavioral Phenotypes for Engagement and Health Outcome Efficacy of an mHealth Intervention for Obesity: Randomized Controlled Trial." In: *Journal of medical Internet research* 23.6 (2021), e27218.

[94] Zachary D King et al. "Micro-stress EMA: A passive sensing framework for detecting in-the-wild stress in pregnant mothers." In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3.3 (2019), pp. 1–22.

[95] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980* (2014).

[96] C Kirschbaum, S Wüst, HG Faig, and DH Hellhammer. "Heritability of cortisol responses to human corticotropin-releasing hormone, ergometry, and psychological stress in humans." In: *The Journal of Clinical Endocrinology & Metabolism* 75.6 (1992), pp. 1526–1530.

[97] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. "The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting." In: *Neuropsychobiology* 28.1-2 (1993), pp. 76–81.

[98] Kent C Kowalski, Peter RE Crocker, and Nanette P Kowalski. "Convergent validity of the physical activity questionnaire for adolescents." In: *Pediatric exercise science* 9.4 (1997), pp. 342–352.

[99] Richard L Kravitz, Naihua Duan, Sunita Vohra, Jiang Li, et al. "Introduction to N-of-1 trials: indications and barriers." In: *Design and Implementation of N-of-1 Trials: A User's Guide* (2014), pp. 1–11.

[100] Sylvia D Kreibig. "Autonomic nervous system activity in emotion: A review." In: *Biological psychology* 84.3 (2010), pp. 394–421.

[101] Kurt Kroenke and Robert L Spitzer. *The PHQ-9: a new depression diagnostic and severity measure*. 2002.

[102] Konstantinos Kyritsis, Christina Lefkothea Tatli, Christos Diou, and Anastasios Delopoulos. "Automated analysis of in meal eating behavior using a commercial wristband IMU sensor." In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2017, pp. 2843–2846.

[103] Michael Lecocke and Kenneth Hess. "An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data." In: *Cancer Informatics* 2 (2006), p. 117693510600200016.

[104] Boon Giin Lee, Jae-Hee Park, Chuan Chin Pu, and Wan-Young Chung. "Smartwatch-based driver vigilance indicator with kernel-fuzzy-C-means-wavelet method." In: *IEEE sensors journal* 16.1 (2015), pp. 242–253.

[105] Dingwen Li, Patrick Lyons, Jeff Klaus, Brian Gage, Marin Kollef, and Chenyang Lu. "Integrating Static and Time-Series Data in Deep Recurrent Models for Oncology Early Warning Systems." In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 913–936.

[106] Dingwen Li, Jay Vaidya, Michael Wang, Ben Bush, Chenyang Lu, Marin Kollef, and Thomas Bailey. "Feasibility Study of Monitoring Deterioration of Outpatients Using Multimodal Data Collected by Wearables." In: *ACM Transactions on Computing for Healthcare* 1.1 (2020), pp. 1–22.

[107] Daniyal Liaqat et al. "WearBreathing: Real World Respiratory Rate Monitoring Using Smartwatches." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.2 (2019), pp. 1–22.

[108] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. "The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?" In: *Personalized medicine* 8.2 (2011), pp. 161–173.

[109] Yue-Der Lin, Ya-Hsueh Chien, and Yi-Sheng Chen. "Wavelet-based embedded algorithm for respiratory rate estimation from PPG signal." In: *Biomedical Signal Processing and Control* 36 (2017), pp. 138–145.

[110] Sally K Longmore, Gough Y Lui, Ganesh Naik, Paul P Breen, Bin Jalaludin, and Gaetano D Gargiulo. "A comparison of reflective photoplethysmography for detection of heart rate, blood oxygen saturation, and respiration rate at various anatomical locations." In: *Sensors* 19.8 (2019), p. 1874.

[111] Louise Lotfi, Lena Flyckt, Ingvar Krakau, Björn Mårtensson, and Gunnar H Nilsson. "Undetected depression in primary healthcare: occurrence, severity and co-morbidity in a two-stage procedure of opportunistic screening." In: *Nordic journal of psychiatry* 64.6 (2010), pp. 421–427.

[112] Carissa A Low, Kristen Salomon, and Karen A Matthews. "Chronic life stress, cardiovascular reactivity, and subclinical cardiovascular disease in adolescents." In: *Psychosomatic medicine* 71.9 (2009), p. 927.

[113] Jin Lu et al. "Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018), pp. 1–21.

[114] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions." In: *Proceedings of the 31st international conference on neural information processing systems*. 2017, pp. 4768–4777.

[115] Nan Lv et al. "The ENGAGE-2 study: Engaging self-regulation targets to understand the mechanisms of behavior change and improve mood and weight outcomes in a randomized controlled trial (Phase 2)." In: *Contemporary Clinical Trials* 95 (2020), p. 106072.

[116] Jun Ma et al. "Reduced Nonconscious Reactivity to Threat in Amygdala Mediates Physical Activity and Energy Expenditure in Integrated Behavior Therapy for Adults with Obesity and Comorbid Depression." In: *CIRCULATION*. Vol. 141. 2020.

[117] Shamona Maharaj, Ty Lees, and Sara Lal. "Prevalence and risk factors of depression, anxiety, and stress in a cohort of Australian nurses." In: *International journal of environmental research and public health* 16.1 (2019), p. 61.

[118] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015.

[119] David J Meredith, D Clifton, P Charlton, J Brooks, CW Pugh, and L Tarassenko. "Photoplethysmographic derivation of respiratory rate: a review of relevant physiology." In: *Journal of medical engineering & technology* 36.1 (2012), pp. 1–7.

[120] RD Mirza, S Punja, S Vohra, and G Guyatt. "The history and development of N-of-1 trials." In: *Journal of the Royal Society of Medicine* 110.8 (2017), pp. 330–340.

[121] Kenneth E Mobily, Linda M Rubenstein, Jon H Lemke, Michael W O'Hara, and Robert B Wallace. "Walking and depression in a cohort of older adults: The Iowa 65+ Rural Health Study." In: *Journal of Aging and Physical Activity* 4.2 (1996), pp. 119–135.

[122] Jennifer N Morey, Ian A Boggero, April B Scott, and Suzanne C Segerstrom. "Current directions in stress and human immune function." In: *Current opinion in psychology* 5 (2015), pp. 13–17.

[123] K Mostov, E Liptsen, and R Boutchko. "Medical applications of shortwave FM radar: Remote monitoring of cardiac and respiratory motion." In: *Medical physics* 37.3 (2010), pp. 1332–1338.

[124] Susan A Murphy and Derek Bingham. "Screening experiments for developing dynamic treatment regimes." In: *Journal of the American Statistical Association* 104.485 (2009), pp. 391–408.

[125] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Jacob E Sunshine. "Opioid overdose detection using smartphones." In: *Science translational medicine* 11.474 (2019).

[126] Shamim Nemati, Atul Malhotra, and Gari D Clifford. "Data fusion for improved respiration rate estimation." In: *EURASIP journal on advances in signal processing* 2010.1 (2010), p. 926305.

[127] Ada Ng, Madhu Reddy, Alyson K Zalta, Stephen M Schueller, et al. "Veterans' perspectives on fitbit use in treatment for post-traumatic stress disorder: an interview study." In: *JMIR mental health* 5.2 (2018), e10415.

[128] NH Department of Administrative. *Services Perceived Stress Scale*. 2020.

[129] PM Nielsen, IJ Le Grice, BH Smaill, and PJ Hunter. "Mathematical model of geometry and fibrous structure of the heart." In: *American Journal of Physiology-Heart and Circulatory Physiology* 260.4 (1991), H1365–H1378.

[130] Rudolf E Noble. "Depression in women." In: *Metabolism* 54.5 (2005), pp. 49–52.

[131] T Christian North, PENNY McCullagh, Zung Vu Tran, David Ed Lavallee, Jean M Williams, Marc V Jones, and Anthony Col Papathomas. "Effect of exercise on depression." In: (2008).

[132] Peter J Norton. "Depression Anxiety and Stress Scales (DASS-21): Psychometric analysis across four racial groups." In: *Anxiety, stress, and coping* 20.3 (2007), pp. 253–265.

[133] Aaron van den Oord et al. "Wavenet: A generative model for raw audio." In: *arXiv preprint arXiv:1609.03499* (2016).

[134] World Health Organization et al. *Depression and other common mental disorders: global health estimates*. Tech. rep. World Health Organization, 2017.

[135] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. "Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring." In: *IEEE journal of biomedical and health informatics* 19.3 (2014), pp. 832–838.

[136] Augustine Osman, Jane L Wong, Courtney L Bagge, Stacey Freedenthal, Peter M Gutierrez, and Gregorio Lozano. "The depression anxiety stress Scales—21 (DASS-21): further examination of dimensions, scale reliability, and correlates." In: *Journal of clinical psychology* 68.12 (2012), pp. 1322–1338.

[137] Anstasopoulou Panagiota, Shammas Layal, and Hey Stefan. "Assessment of human gait speed and energy expenditure using a single triaxial accelerometer." In: *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. IEEE. 2012, pp. 184–188.

[138] Junbiao Pang, Qingming Huang, and Shuqiang Jiang. "Multiple instance boost using graph embedding based decision stump for pedestrian detection." In: *European conference on computer vision*. Springer. 2008, pp. 541–552.

[139] David Paper and David Paper. "Scikit-Learn Classifier Tuning from Simple Training Sets." In: *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python* (2020), pp. 137–163.

[140] Jaeyeon Park, Woojin Nam, Jaewon Choi, Taeyeong Kim, Dukyong Yoon, Sukhoon Lee, Jeongyeup Paek, and JeongGil Ko. "Glasses for the third eye: Improving the quality of clinical data analysis with motion sensor-based data filtering." In: *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 2017, pp. 1–14.

[141] Onur Parlak, Scott Tom Keene, Andrew Marais, Vincenzo F Curto, and Alberto Salleo. "Molecularly selective nanoporous membrane-based wearable organic electrochemical device for noninvasive cortisol sensing." In: *Science advances* 4.7 (2018), eaar2904.

[142] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035.

[143] Meenal J Patel, Alexander Khalaf, and Howard J Aizenstein. "Studying depression using imaging and machine learning methods." In: *NeuroImage: Clinical* 10 (2016), pp. 115–123.

[144] Neal Patwari, Joey Wilson, Sai Ananthanarayanan, Sneha K Kasera, and Dwayne R Westenskow. "Monitoring breathing via signal strength in wireless networks." In: *IEEE Transactions on Mobile Computing* 13.8 (2013), pp. 1774–1786.

[145] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[146] Thomas Penzel, Jan W Kantelhardt, Ludger Grote, Jörg-Hermann Peter, and Armin Bunde. "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea." In: *IEEE Transactions on biomedical engineering* 50.10 (2003), pp. 1143–1151.

[147] Marisa J Perera et al. "Factor structure of the Perceived Stress Scale-10 (PSS) across English and Spanish language responders in the HCHS/SOL Sociocultural Ancillary Study." In: *Psychological assessment* 29.3 (2017), p. 320.

[148] Marco V Perez et al. "Large-scale assessment of a smartwatch to identify atrial fibrillation." In: *New England Journal of Medicine* 381.20 (2019), pp. 1909–1917.

[149] Marco AF Pimentel, Alistair EW Johnson, Peter H Charlton, Drew Birrenkott, Peter J Watkinson, Lionel Tarassenko, and David A Clifton. "Toward a robust estimation of respiratory rate from pulse oximeters." In: *IEEE Transactions on Biomedical Engineering* 64.8 (2016), pp. 1914–1923.

[150] Marco AF Pimentel, Alistair EW Johnson, Peter H Charlton, Drew Birrenkott, Peter J Watkinson, Lionel Tarassenko, and David A Clifton. "Toward a robust estimation of respiratory rate from pulse oximeters." In: *IEEE Transactions on Biomedical Engineering* 64.8 (2016), pp. 1914–1923.

[151] Kurt Plarre et al. "Continuous inference of psychological stress from sensory measurements collected in the natural environment." In: *Proceedings of the 10th ACM/IEEE international conference on information processing in sensor networks*. IEEE. 2011, pp. 97–108.

[152] Rüdiger Pryss et al. "Exploring the time trend of stress levels while using the crowdsensing mobile health platform, trackyourstress, and the influence of perceived stress reactivity: ecological momentary assessment pilot study." In: *JMIR mHealth and uHealth* 7.10 (2019), e13978.

[153] *PyTorch Mobile*. 2020. URL: `https://pytorch.org/mobile/home/`.

[154] Hude Quan et al. "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data." In: *Medical care* (2005), pp. 1130–1139.

[155] Elizabeth Quattrocki, Abigail Baird, and Deborah Yurgelun-Todd. "Biological aspects of the link between smoking and depression." In: *Harvard review of psychiatry* 8.3 (2000), pp. 99–110.

[156] Antoine Raberin et al. "Role of gender and physical activity level on cardiovascular risk factors and biomarkers of oxidative stress in the elderly." In: *Oxidative Medicine and Cellular Longevity* 2020 (2020).

[157] Valentin Radu, Nicholas D Lane, Sourav Bhattacharya, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. "Towards multimodal deep learning for activity recognition on mobile devices." In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016, pp. 185–188.

[158] Alfredo Raglio et al. "Machine learning techniques to predict the effectiveness of music therapy: A randomized controlled trial." In: *Computer methods and programs in biomedicine* 185 (2020), p. 105160.

[159] Piyush Rai, Abhishek Kumar, and Hal Daume. "Simultaneously leveraging output and task structures for multiple-output regression." In: *Advances in Neural Information Processing Systems* 25 (2012), pp. 3185–3193.

[160] Satu Rajala, Harri Lindholm, and Tapio Taipalus. "Comparison of photoplethysmogram measured from wrist and finger and the effect of measurement location on pulse arrival time." In: *Physiological measurement* 39.7 (2018), p. 075010.

[161] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet. "SimpleMKL." In: *Journal of Machine Learning Research* 9 (2008), pp. 2491–2521.

[162] M Raghu Ram, K Venu Madhav, E Hari Krishna, Nagarjuna Reddy Komalla, and K Ashoka Reddy. "A novel approach for motion artifact reduction in PPG signals based on AS-LMS adaptive filter." In: *IEEE Transactions on Instrumentation and Measurement* 61.5 (2011), pp. 1445–1457.

[163] Vignesh Ravichandran, Balamurali Murugesan, Vaishali Balakarthikeyan, Keerthi Ram, SP Preejith, Jayaraj Joseph, and Mohanasankar Sivaprakasam. "RespNet: A deep learning model for extraction of respiration from photoplethysmogram." In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2019, pp. 5556–5559.

[164] Amanda L Rebar, Robert Stanton, David Geard, Camille Short, Mitch J Duncan, and Corneel Vandelanotte. "A meta-meta-analysis of the effect of physical activity on depression and anxiety in non-clinical adult populations." In: *Health psychology review* 9.3 (2015), pp. 366–378.

[165] Andrew Reisner, Phillip A Shaltis, Devin McCombie, and H Harry Asada. "Utility of the photoplethysmogram in circulatory monitoring." In: *Anesthesiology: The Journal of the American Society of Anesthesiologists* 108.5 (2008), pp. 950–958.

[166] Eduardo Remor. "Psychometric properties of a European Spanish version of the Perceived Stress Scale (PSS)." In: *The Spanish journal of psychology* 9.1 (2006), pp. 86–93.

[167] Bersain A Reyes, Natasa Reljin, Youngsun Kong, Yunyoung Nam, and Ki H Chon. "Tidal volume and instantaneous respiration rate estimation using a volumetric surrogate signal acquired via a smartphone camera." In: *IEEE journal of biomedical and health informatics* 21.3 (2016), pp. 764–777.

[168] Antônio H Ribeiro, Koen Tiels, Luis A Aguirre, and Thomas Schön. "Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2370–2380.

[169] Ícaro JS Ribeiro, Rafael Pereira, Ivna V Freire, Bruno G de Oliveira, Cezar A Casotti, and Eduardo N Boery. "Stress and quality of life among university students: A systematic literature review." In: *Health Professions Education* 4.2 (2018), pp. 70–77.

[170] Bryan Rodgers, AE Korten, AF Jorm, PA Jacomb, H Christensen, and AS Henderson. "Non-linear relationships in associations of depression and anxiety with alcohol use." In: *Psychological medicine* 30.2 (2000), pp. 421–432.

[171] Douglas K Russell. "The Boltzmann distribution." In: *Journal of Chemical Education* 73.4 (1996), p. 299.

[172] Yvan Saeys, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics." In: *bioinformatics* 23.19 (2007), pp. 2507–2517.

[173] Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. "A weakly supervised learning framework for detecting social anxiety and depression." In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2.2 (2018), pp. 1–26.

[174] Robert J Schalkoff. "Pattern recognition." In: *Wiley Encyclopedia of Computer Science and Engineering* (2007).

[175] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. "Introducing wesad, a multimodal dataset for wearable stress and affect detection." In: *Proceedings of the 20th ACM international conference on multimodal interaction*. 2018, pp. 400–408.

[176] Suzanne C Segerstrom and Gregory E Miller. "Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry." In: *Psychological bulletin* 130.4 (2004), p. 601.

[177] Hans Selye. "Stress without distress." In: *Psychopathology of human adaptation*. Springer, 1976, pp. 137–146.

[178] Umit Sener, Kagan Ucok, Alper M Ulasli, Abdurrahman Genc, Hatice Karabacak, Necip F Coban, Hasan Simsek, and Halime Cevik. "Evaluation of health-related physical fitness parameters and association analysis with depression, anxiety, and quality of life in patients with fibromyalgia." In: *International Journal of Rheumatic Diseases* 19.8 (2016), pp. 763–772.

[179] Fernando Seoane, Inmaculada Mohino-Herranz, Javier Ferreira, Lorena Alvarez, Ruben Buendia, David Ayllón, Cosme Llerena, and Roberto Gil-Pita. "Wearable biomedical measurement systems for assessment of mental stress of combatants in real time." In: *Sensors* 14.4 (2014), pp. 7120–7141.

[180] Fred Shaffer and Jay P Ginsberg. "An overview of heart rate variability metrics and norms." In: *Frontiers in public health* (2017), p. 258.

[181] Yichen Shen, Maxime Voisin, Alireza Aliamiri, Anand Avati, Awni Hannun, and Andrew Ng. "Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning." In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 1909–1916.

[182] Saul Shiffman, Arthur A Stone, and Michael R Hufford. "Ecological momentary assessment." In: *Annu. Rev. Clin. Psychol.* 4 (2008), pp. 1–32.

[183] Hang Sik Shin, Chungkeun Lee, and Myoungho Lee. "Adaptive threshold method for the peak detection of photoplethysmographic waveform." In: *Computers in biology and medicine* 39.12 (2009), pp. 1145–1152.

[184] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3145–3153.

[185] Rajita Sinha. "Chronic stress, drug use, and vulnerability to addiction." In: *Annals of the new York Academy of Sciences* 1141.1 (2008), pp. 105–130.

[186] Rodrigo Siqueira Reis, Adriano Akira Ferreira Hino, and Ciro Romélio Rodriguez Añez. "Perceived stress scale: reliability and validity study in Brazil." In: *Journal of health psychology* 15.1 (2010), pp. 107–114.

[187] Elena Smets, Walter De Raedt, and Chris Van Hoof. "Into the wild: the challenges of physiological stress detection in laboratory and ambulatory settings." In: *IEEE journal of biomedical and health informatics* 23.2 (2018), pp. 463–473.

[188] Elena Smets, Giuseppina Schiavone, Emmanuel Rios Velazquez, Walter De Raedt, Katleen Bogaerts, Ilse Van Diest, and Chris Van Hoof. "Comparing task-induced psychophysiological responses between persons with stress-related complaints and healthy controls: A methodological pilot study." In: *Health Science Reports* 1.8 (2018), e60.

[189] Elena Smets et al. "Large-scale wearable data reveal digital phenotypes for daily-life stress detection." In: *NPJ digital medicine* 1.1 (2018), pp. 1–10.

[190] Julius Orion Smith. *Introduction to digital filters: with audio applications*. Vol. 2. Julius Smith, 2007.

[191] Megan V Smith and Carolyn M Mazure. "Mental health and wealth: depression, gender, poverty, and parenting." In: *Annual review of clinical psychology* 17 (2021), pp. 181–205.

[192] Matthew Smuck, Charles A Odonkor, Jonathan K Wilt, Nicolas Schmidt, and Michael A Swiernik. "The emerging clinical role of wearables: factors for successful implementation in healthcare." In: *NPJ Digital Medicine* 4.1 (2021), pp. 1–8.

[193] Marika B Solhan, Timothy J Trull, Seungmin Jahng, and Phillip K Wood. "Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall." In: *Psychological assessment* 21.3 (2009), p. 425.

[194] Kent A Spackman, Keith E Campbell, and Roger A Côté. "SNOMED RT: a reference terminology for health care." In: *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association. 1997, p. 640.

[195] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. "A brief measure for assessing generalized anxiety disorder: the GAD-7." In: *Archives of internal medicine* 166.10 (2006), pp. 1092–1097.

[196] Andreas Ströhle. "Physical activity, exercise, depression and anxiety disorders." In: *Journal of neural transmission* 116.6 (2009), pp. 777–784.

[197] Xiao Sun, Li Qiu, Yibo Wu, Yeming Tang, and Guohong Cao. "Sleepmonitor: Monitoring respiratory rate and body position during sleep using smartwatch." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–22.

[198] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[199] John WG Tiller. "Depression and anxiety." In: *The Medical Journal of Australia* 199.6 (2013), S28–S31.

[200] John Trimpop, Hannes Schenk, Gerald Bieber, Friedrich Lämmel, and Paul Burggraf. "Smartwatch based respiratory rate and breathing pattern recognition in an end-consumer environment." In: *Proceedings of the 4th international Workshop on Sensor-based Activity Recognition and Interaction*. 2017, pp. 1–5.

[201] Diane M Turner-Bowker, Martha S Bayliss, John E Ware, and Mark Kosinski. "Usefulness of the SF-8™ Health Survey for comparing the impact of migraine and other conditions." In: *Quality of Life Research* 12.8 (2003), pp. 1003–1012.

[202] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks." In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 2017, pp. 216–220.

[203] All of Us Research Program Investigators. "The "All of Us" research program." In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676.

[204] HM Van Praag. "Can stress cause depression?" In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 28.5 (2004), pp. 891–907.

[205] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In: *Advances in neural information processing systems* 30 (2017).

[206] L Ventzel, CS Madsen, AB Jensen, AR Jensen, TS Jensen, and NB Finnerup. "Assessment of acute oxaliplatin-induced cold allodynia: a pilot study." In: *Acta Neurologica Scandinavica* 133.2 (2016), pp. 152–155.

[207] John Wallert, Emelie Gustafson, Claes Held, Guy Madison, Fredrika Norlund, Louise von Essen, and Erik Martin Gustaf Olsson. "Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial." In: *Journal of medical Internet research* 20.10 (2018), e10754.

[208] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. "Tracking depression dynamics in college students using mobile phone and wearable sensing." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018), pp. 1–26.

[209] Rui Wang et al. "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones." In: *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 2014, pp. 3–14.

[210] Siying Wang, Antje Pohl, Timo Jaeschke, Michael Czaplik, Marcus Köny, Steffen Leonhardt, and Nils Pohl. "A novel ultra-wideband 80 GHz FMCW radar system for contactless monitoring of vital signs." In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 4978–4981.

[211] Xuyu Wang, Chao Yang, and Shiwen Mao. "TensorBeat: Tensor decomposition for monitoring multiperson breathing beats with commodity WiFi." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 9.1 (2017), pp. 1–27.

[212] Shweta Ware et al. "Large-scale automatic depression screening using meta-data from wifi infrastructure." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.4 (2018), pp. 1–27.

[213] *Wearable Medical Devices Market Size, Share & Growth [2027].* `https://www.fortunebusinessinsights.com/industry-reports/wearable-medical-devices-market-101070`. (Accessed on 05/16/2021).

[214] *What are Active Zone Minutes or active minutes on my Fitbit device?* `https://help.fitbit.com/articles/en_US/Help_article/1379.htm`. (Accessed on 10/06/2021).

[215] Leanne M Williams et al. "Sensitivity, specificity, and predictive power of the "Brief Risk-resilience Index for Screening," a brief pan-diagnostic web screen for emotional health." In: *Brain and behavior* 2.5 (2012), pp. 576–589.

[216] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. "Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals." In: *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 2019, pp. 54–66.

[217] Xuhai Xu et al. "Leveraging routine behavior and contextually-filtered features for depression detection among college students." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.3 (2019), pp. 1–33.

[218] Lan Yu, Daniel J Buysse, Anne Germain, Douglas E Moul, Angela Stover, Nathan E Dodds, Kelly L Johnston, and Paul A Pilkonis. "Development of short forms from the PROMIS™ sleep disturbance and sleep-related impairment item banks." In: *Behavioral sleep medicine* 10.1 (2012), pp. 6–24.

[219] Mari Zakrzewski, Antti Vehkaoja, Atte S Joutsen, Karri T Palovuori, and Jukka J Vanhala. "Noncontact respiration monitoring during sleep with microwave Doppler radar." In: *IEEE Sensors Journal* 15.10 (2015), pp. 5683–5693.

[220] Fuquan Zhang et al. "Genetic evidence suggests posttraumatic stress disorder as a subtype of major depressive disorder." In: *The Journal of clinical investigation* 132.3 (2022).

[221] Yuezhou Zhang et al. "Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study." In: *JMIR mHealth and uHealth* 9.4 (2021), e24604.

[222] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. "Informer: Beyond efficient transformer for long sequence time-series forecasting." In: *Proceedings of AAAI*. 2021.