Multi-Task Learning for Randomized Controlled Trials: A Case Study on Predicting Depression with Wearable Data

RUIXUAN DAI, Washington University in St. Louis, United States
THOMAS KANNAMPALLIL, Washington University in St. Louis, United States
JINGWEN ZHANG, Washington University in St. Louis, United States
NAN LV, University of Illinois Chicago, United States
JUN MA, University of Illinois Chicago, United States
CHENYANG LU, Washington University in St. Louis, United States

A randomized controlled trial (RCT) is used to study the safety and efficacy of new treatments, by comparing patient outcomes of an intervention group with a control group. Traditionally, RCTs rely on statistical analyses to assess the differences between the treatment and control groups. However, such statistical analyses are generally not designed to assess the impact of the intervention at an individual level. In this paper, we explore machine learning models in conjunction with an RCT for personalized predictions of a depression treatment intervention, where patients were longitudinally monitored with wearable devices. We formulate individual-level predictions in the intervention and control groups from an RCT as a multi-task learning (MTL) problem, and propose a novel MTL model specifically designed for RCTs. Instead of training separate models for the intervention and control groups, the proposed MTL model is trained on both groups, effectively enlarging the training dataset. We develop a hierarchical model architecture to aggregate data from different sources and different longitudinal stages of the trial, which allows the MTL model to exploit the commonalities and capture the differences between the two groups. We evaluated the MTL approach in an RCT involving 106 patients with depression, who were randomized to receive an integrated intervention treatment. Our proposed MTL model outperforms both single-task models and the traditional multi-task model in predictive performance, representing a promising step in utilizing data collected in RCTs to develop predictive models for precision medicine.

CCS Concepts: • **Applied computing** → **Consumer health**; • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: Wearables, Depression, Randomized Controlled Trial, Multi-task Learning

ACM Reference Format:
Ruixuan Dai, Thomas Kannampallil, Jingwen Zhang, Nan Lv, Jun Ma, and Chenyang Lu. 2022. Multi-Task Learning for Randomized Controlled Trials: A Case Study on Predicting Depression with Wearable Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 50 (June 2022), 23 pages. https://doi.org/10.1145/3534591

Authors' addresses: Ruixuan Dai, Washington University in St. Louis, Department of Computer Science and Engineering, 1 Brookings Dr., St. Louis, Missouri, 63130, United States, dairuixuan@wustl.edu; Thomas Kannampallil, Washington University in St. Louis, Department of Anesthesiology, Department of Computer Science and Engineering, 1 Brookings Dr., St. Louis, Missouri, 63130, United States, thomas.k@wustl.edu; Jingwen Zhang, Washington University in St. Louis, Department of Computer Science and Engineering, 1 Brookings Dr., St. Louis, Missouri, 63130, United States, jingwen.z@wustl.edu; Nan Lv, University of Illinois Chicago, Department of Medicine, 1200 West Harrison St., Chicago, Illinois, 60607, United States, lvn2017@uic.edu; Jun Ma, University of Illinois Chicago, Department of Medicine, 1200 West Harrison St., Chicago, Illinois, 60607, United States, maj2015@uic.edu; Chenyang Lu, Washington University in St. Louis, Department of Computer Science and Engineering, 1 Brookings Dr., St. Louis, Missouri, 63130, United States, lu@wustl.edu.

2474-9567/2022/6-ART50
https://doi.org/10.1145/3534591

1 INTRODUCTION

A *randomized controlled trial (RCT)* is considered the gold standard for evaluating treatment efficacy, including in the case of mental health interventions [26]. 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 [6].

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 [58]. This "silent killer" is a major public health burden costing more than $1 trillion US dollars every year [28]. Wearable devices provide a convenient way for continuous remote activity monitoring, owing to their popularity, pervasive availability and relatively low cost. Recent studies [11, 56, 75, 76] have been reasonably successful in tracking depression using wearables, showing the association between depression and physical activity [57, 64]. 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* [39]. If the model is developed on the intervention group to predict the intervention outcomes, it is called *treatment-specific modeling* [39]. 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 [26]; (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.

2 RELATED WORK

2.1 Mental Health with Mobile and Wearable Devices

Mental health disorders, such as depression, anxiety and stress, usually have common attributes [19, 72], and can have an adverse impact on our daily life [18]. 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 [5, 7, 22, 75] 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 [75]. 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 [22]. Voice data from smartphones can also be applied to discover potential markers of mental illnesses using machine learning and natural language processing techniques [14]. 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[31]. Zhang et al. employed a wristband tracker to monitor sleep, and associated the depressive symptom severity with the sleep quality [79]. 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 [36]. 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% [40]. 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 [68].

Recently, several studies [35, 49] 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 [19, 72], MTL could potentially benefit different mental health prediction tasks when learning together. In [49], 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 [35], 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 [49].

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 [6].

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 [10, 41, 62, 74], to determine individual-level predictions. Previous wearable studies usually belong to the risk modeling category [39] 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. [10] 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% [10], 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% [6, 33, 61]. Lhmig et al. [32] 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 [62], 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. [74] 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 [10, 41, 62, 74] focus on either treatment-specific modeling or risk modeling.

Adaptive trials [3, 23, 29] 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, 55]. 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 [23, 44, 48, 53], 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 [44]. 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.

3 CLINICAL TRIAL AND DATA PROCESSING

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

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 [45] 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 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 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[1]) 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 [51]. The bottom part of Figure 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 [12, 13]. 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.

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

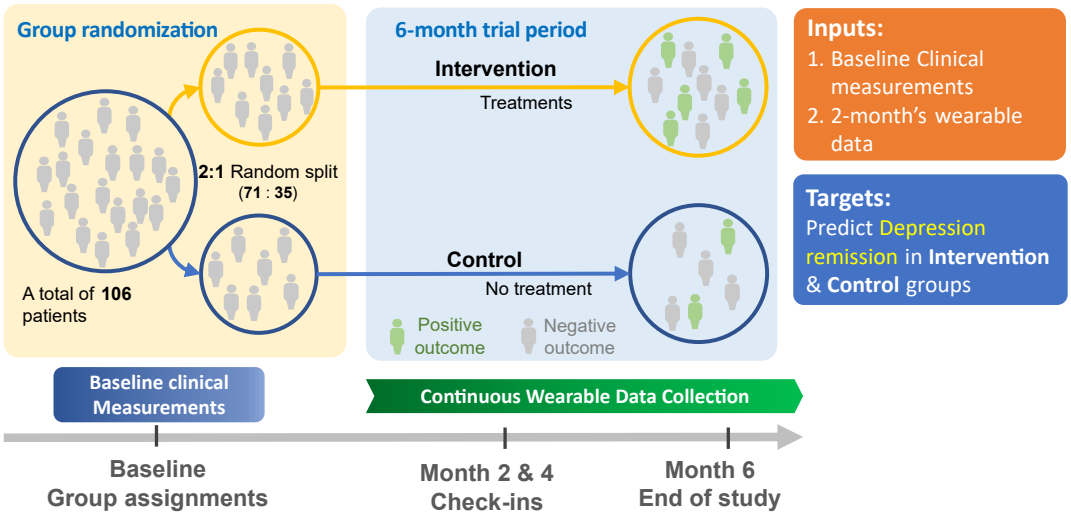


Fig. 1. Diagram of our RCT study.

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 [47]. 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.

3.2.1 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 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.

3.2.2 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.

Table 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 [43]	654.58 ± 819.72	953.20 ± 1131.43	0.158
Work, MET mins/week [43]	336.27 ± 937.05	217.33 ± 999.95	0.581
Energy expenditure, kcal/kg/d [52]	33.52 ± 2.23	33.64 ± 2.32	0.807
SPSI-R:S raw score [16]	12.88 ± 2.41	13.13 ± 2.40	0.639
PPO raw score [16]	10.59 ± 4.54	12.58 ± 3.33	0.037
NPO raw score [16]	7.07 ± 3.48	8.62 ± 3.52	0.051
RPS raw score [16]	10.46 ± 4.48	11.29 ± 3.67	0.382
ICS raw score [16]	4.05 ± 3.65	4.74 ± 3.22	0.382
AS raw score [16]	5.54 ± 4.54	4.84 ± 3.46	0.461
PROMIS sleep disturbance t score [78]	57.96 ± 7.44	57.08 ± 8.05	0.611
PROMIS sleep impairment t score [78]	56.95 ± 8.80	54.83 ± 9.09	0.291
SCL-20 score [24]	1.21 ± 0.67	1.15 ± 0.59	0.695
GAD-7 score [70]	7.10 ± 5.12	6.63 ± 3.77	0.659
PTSD severity score [4]	36.12 ± 14.57	33.77 ± 10.25	0.432
SF-8 physical component score [73]	45.00 ± 8.48	47.38 ± 8.30	0.21
SF-8 mental component score [73]	39.65 ± 11.37	42.57 ± 9.00	0.225
COPE total scores [9]	57.02 ± 12.09	57.40 ± 14.63	0.896
BRISC total scores [77]	30.12 ± 6.23	31.07 ± 6.06	0.495

Abbreviations: AS, avoidance style; COPE, COPE Inventory survey, including 14 components [9]; BRISC, BRISC questionnaire of emotional resilience and self-efficacy Survey, including 3 components [77]; 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.

- **Sedentary minutes (sedentaryMinutes).** Fitbit provides estimations of the active minutes through the metabolic equivalents (METs) [1, 37]. 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 [8]. 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) [30] 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 [25]. 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.

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 [49, 76]. 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" [67]. As such, we employed a high-level feature engineering approach [47] 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.

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 [17] 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 [17, 27, 66]. 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 [46].

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) [35], 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.

4.1 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 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.

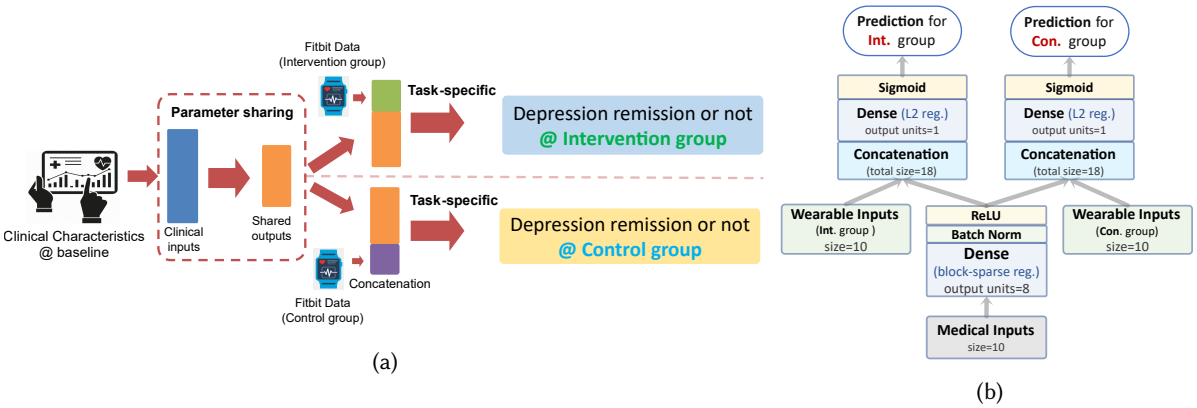


Fig. 2. (a) MTL framework for randomized controlled trials; (b) MTL Model structure.

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 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 [34, 63], 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 [34], 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 [63], 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:

$$R_{blk} = \| |a_1|_2, |a_2|_2, ..., |a_n|_2 \|_1 \tag{1}$$

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:

$$R_{tsk,j} = \left\|b_j\right\|_2 \tag{2}$$

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 3 shows the overall masked classification loss function for all tasks:

$$Loss_{clf} = \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} w_t \cdot mask_{n,t} \cdot BCE(y_{n,t}, \hat{y}_{n,t}) \tag{3}$$

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.

4.3 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 [38]. Previous work [38] 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 [65], 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:

$$y'_{n,t} = p(y_{n,t} = 1 | f_t(x_n), \sigma_t) = Sigmoid(\frac{1}{\sigma_t^2} f_t(x_n)) = \frac{e^{f_t(x_n)/\sigma_t^2}}{1 + e^{f_t(x_n)/\sigma_t^2}} \tag{4}$$

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 3, we have:

$$Loss'_{clf} = \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} mask_{n,t} \cdot BCE(y'_{n,t}, \hat{y}_{n,t}) \tag{5}$$

$$= \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} mask_{n,t} \cdot [-\hat{y}_{n,t} \cdot log(y'_{n,t}) - (1 - \hat{y}_{n,t}) \cdot log(1 - y'_{n,t})] \tag{6}$$

$$= \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} mask_{n,t} \cdot [-\hat{y}_{n,t} \frac{f_t(x_n)}{\sigma_t^2} + log(1 + e^{f_t(x_n)/\sigma_t^2})] \tag{7}$$

$$= \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} mask_{n,t} \cdot [-\frac{1}{\sigma_t^2} \hat{y}_{n,t} f_t(x_n) + \frac{1}{\sigma_t^2} log(1 + e^{f_t(x_n)}) + log \frac{1 + e^{f_t(x_n)/\sigma_t^2}}{(1 + e^{f_t(x_n)})^{\frac{1}{\sigma_t^2}}}] \tag{8}$$

$$= \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} mask_{n,t} \cdot [\frac{1}{\sigma_t^2} BCE(y_{n,t}, \hat{y}_{n,t}) + log \frac{1 + e^{f_t(x_n)/\sigma_t^2}}{(1 + e^{f_t(x_n)})^{\frac{1}{\sigma_t^2}}}] \tag{9}$$

$$\approx \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} \frac{1}{\sigma_t^2} \cdot mask_{n,t} \cdot BCE(y_{n,t}, \hat{y}_{n,t}) + \sum_{t=1}^{T} \log(\sigma_t) \tag{10}$$

In Equation (9), we used an approximation: $(1 + e^{f_t(x_n)})^{1/\sigma_t^2} \approx 1/\sigma_t^2 (1 + e^{f_t(x)/\sigma_t^2})$ [38]. When $\sigma_t = 1$, it becomes equality. In Equation (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 (3) with Equation (10) and summing up the regularization terms, we can then obtain the updated total loss for our MTL framework:

$$Loss = \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} \frac{1}{\sigma_t^2} \cdot mask_{n,t} \cdot BCE(y_{i,t}, \hat{y}_{i,t}) + \sum_{t=1}^{T} \log(\sigma_t) + R_{blk} + \sum_{t=1}^{T} R_{tsk,t} \tag{11}$$

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.

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.

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 [2]. We used an Adam optimizer (learning_rate=0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) [42], 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 [60] to achieve the best AUROC. Table 2 lists all the hyperparameters tuned in our experiment.

Table 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 [59]
	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			

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 3 shows all the selected features and their corresponding selected times.

5.3 MTL vs. STL

Table 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 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

Table 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.

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.

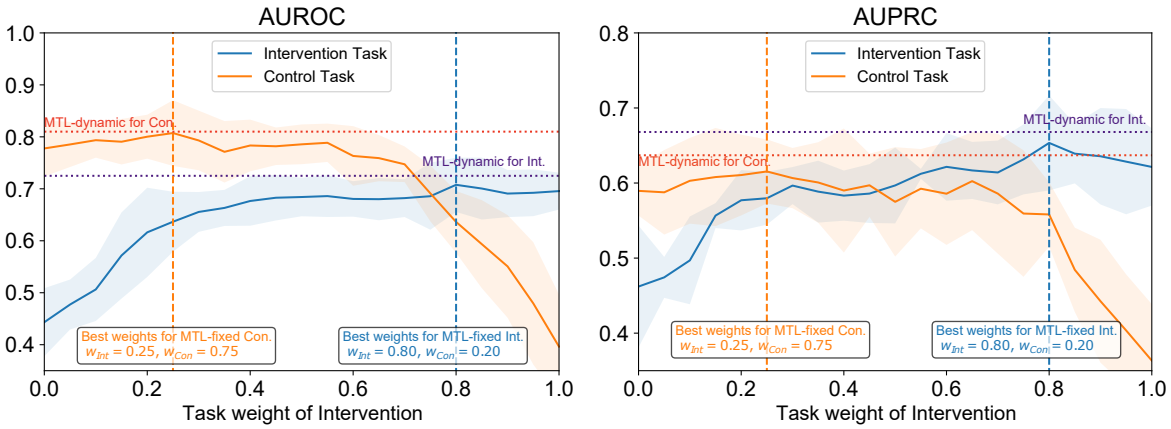
When comparing our MTL models with the best STL models, we observe that our MTL-1 models show comparable performance to the best STL models. The MTL-1 is effectively single-task learning, since it only utilized the information from one group. There is no performance gain from the MTL-1 model, suggesting 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 [20], $p < 0.05$). The MTL-fixed and MTL-dynamic models were trained on the combined 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 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 $w_{Int} = 0.85$

Table 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.



(a) AUROC for the two tasks

(b) AUPRC for the two tasks

Fig. 3. Performances with varying task weights.

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 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 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 [38].

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 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 [49, 76] also demonstrated that wearable data show predictive power in mental health applications.

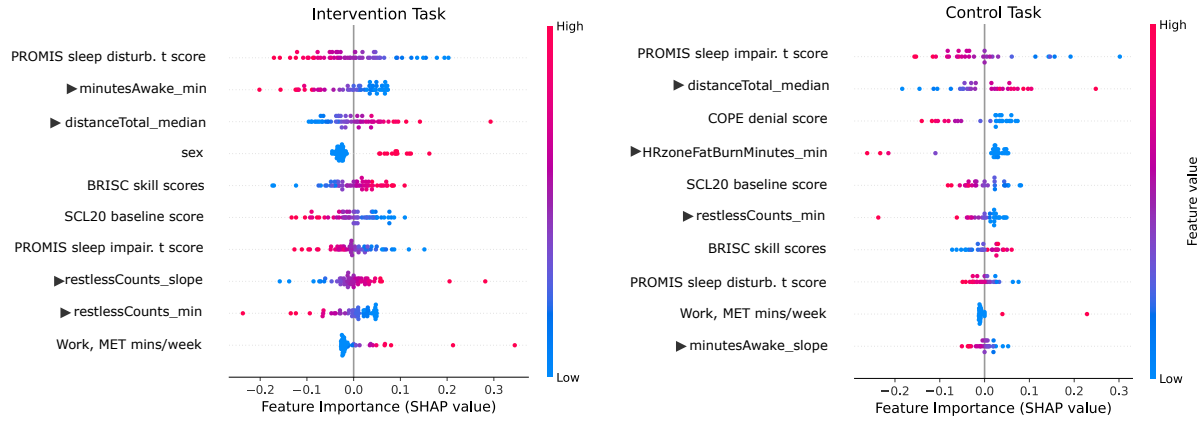
Table 5. Performance comparison with and without wearable data

	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)

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) [50], 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 [50]. The SHAP for deep models is built on a connection with DeepLIFT [69], using a distribution of background samples and Shapley equations to linearize the components in the deep network. Figure 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).



(a) Shap summary plot for the intervention task

(b) Shap summary plot for the control task

Fig. 4. Model explanation for MTL-dynamic models.

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 (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) [77]. 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 [21]. 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 [71] has shown that exercise could be a helpful treatment to depression, and daily walking predicts an improved depression outcome[15, 54]. Our model demonstrates a similar trend as well.

6 DISCUSSION 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 [49], automatic estimations of the outcome could be beneficial to monitoring depression status over time, and potentially assist the doctor in devising personalized treatments. Table 5 and Figure 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 several directions to advance the work presented 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 at a finer granularity. For example, we may develop and train a model to estimate whether a new patient should receive the treatment or not, when we only utilize the baseline data at the first visit.

ACKNOWLEDGMENTS

This research was supported in part by the Fullgraf Foundation; and in part by the National Institutes of Health (NIH) Science of Behavior Change Common Fund Program through an award administered by the National Heart, Lung, and Blood Institute grant number UH2HL132368 (JM, LMW) and UH3HL132368 (JM, LMW). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Dr. Ma is a paid scientific advisor for Health Mentor, Inc. (San Jose, CA, USA). We wish to thank our Data and Safety Monitoring Board members who received compensation for their time: Manisha Desai, Ph.D. (chair), Sandra Tsai, MD, Mickey Trockel, MD, and Manpreet Singh, MD, MS (all from Stanford University). The DSMB reviewed the study protocol initially and data quality and safety monitoring reports semiannually thereafter. We extend special thanks to the participants and their families who made this study possible.

REFERENCES

[1] [n.d.]. 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).

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[3] Daniel Almirall, Inbal Nahum-Shani, Nancy E Sherwood, and Susan A Murphy. 2014. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational behavioral medicine* 4, 3 (2014), 260–274.

[4] Michael A Andrykowski, Matthew J Cordova, Jamie L Studts, and Thomas W Miller. 1998. 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. *Journal of consulting and clinical psychology* 66, 3 (1998), 586.

[5] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal* 38, 3 (2015), 218.

[6] Danilo Bzdok and Andreas Meyer-Lindenberg. 2018. Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3, 3 (2018), 223–230.

[7] Luca Canzian and Mirco Musolesi. 2015. 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*. 1293–1304.

[8] Daniel G Carey. 2009. Quantifying differences in the "fat burning" zone and the aerobic zone: implications for training. *The Journal of Strength & Conditioning Research* 23, 7 (2009), 2090–2095.

[9] Charles S Carver, Michael F Scheier, and Jagdish K Weintraub. 1989. Assessing coping strategies: a theoretically based approach. *Journal of personality and social psychology* 56, 2 (1989), 267.

[10] Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K Johnson, Madhukar H Trivedi, Tyrone D Cannon, John Harrison Krystal, and Philip Robert Corlett. 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry* 3, 3 (2016), 243–250.

[11] Jenny Chum, Min Suk Kim, Laura Zielinski, Meha Bhatt, Douglas Chung, Sharon Yeung, Kathryn Litke, Kathleen McCabe, Jeff Whattam, Laura Garrick, et al. 2017. Acceptability of the Fitbit in behavioural activation therapy for depression: a qualitative study. *Evidence-based mental health* 20, 4 (2017), 128–133.

[12] Paul Ciechanowski, Naomi Chaytor, John Miller, Robert Fraser, Joan Russo, Jurgen Unutzer, and Frank Gilliam. 2010. PEARLS depression treatment for individuals with epilepsy: a randomized controlled trial. *Epilepsy & Behavior* 19, 3 (2010), 225–231.

[13] Paul Ciechanowski, Edward Wagner, Karen Schmaling, Sheryl Schwartz, Barbara Williams, Paula Diehr, Jayne Kulzer, Shelly Gray, Cheza Collier, and James LoGerfo. 2004. Community-integrated home-based depression treatment in older adults: a randomized controlled trial. *Jama* 291, 13 (2004), 1569–1577.

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

[15] F Dimeo, M Bauer, I Varahram, G Proest, and U Halter. 2001. Benefits from aerobic exercise in patients with major depression: a pilot study. *British journal of sports medicine* 35, 2 (2001), 114–117.

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

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

[18] Sherrill Evans, Sube Banerjee, Morven Leese, and Peter Huxley. 2007. The impact of mental illness on quality of life: A comparison of severe mental illness, common mental disorder and healthy population samples. *Quality of life research* 16, 1 (2007), 17–29.

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

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

[21] Julio Fernandez-Mendoza, Sarah Shea, Alexandros N Vgontzas, Susan L Calhoun, Duanping Liao, and Edward O Bixler. 2015. Insomnia and incident depression: role of objective sleep duration and natural history. *Journal of sleep research* 24, 4 (2015), 390–398.

[22] Mads Frost, Gabriela Marcu, Rene Hansen, Karoly Szaántó, and Jakob E Bardram. 2011. 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, 204–205.

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

[24] Richard M Glass, Andrew T Allan, EH Uhlenhuth, Chase P Kimball, and Dennis I Borinstein. 1978. Psychiatric screening in a medical clinic: An evaluation of a self-report inventory. *Archives of General Psychiatry* 35, 10 (1978), 1189–1195.
[25] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. 2019. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. *Journal of medical Internet research* 21, 11 (2019), e16273.
[26] Eduardo Hariton and Joseph J Locascio. 2018. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology* 125, 13 (2018), 1716.
[27] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one* 6, 12 (2011), e28210.
[28] The Lancet Global Health. 2020. Mental health matters. *The Lancet. Global Health* 8, 11 (2020), e1352.
[29] Emily T Hébert, Chaelin K Ra, Adam C Alexander, Angela Helt, Rachel Moisiuc, Darla E Kendzor, Damon J Vidrine, Rachel K Funk-Lawler, and Michael S Businelle. 2020. A mobile Just-in-Time adaptive intervention for smoking cessation: pilot randomized controlled trial. *Journal of medical Internet research* 22, 3 (2020).
[30] CJK Henry. 2005. Basal metabolic rate studies in humans: measurement and development of new equations. *Public health nutrition* 8, 7a (2005), 1133–1152.
[31] Hugh Hunkin, Daniel L King, and Ian T Zajac. 2020. Perceived acceptability of wearable devices for the treatment of mental health problems. *Journal of clinical psychology* 76, 6 (2020), 987–1003.
[32] Frank R Ihmig, Frank Neurohr-Parakenings, Sarah K Schäfer, Johanna Lass-Hennemann, and Tanja Michael. 2020. On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals. *Plos one* 15, 6 (2020), e0231517.
[33] Dan V Iosifescu, Scott Greenwald, Philip Devlin, David Mischoulon, John W Denninger, Jonathan E Alpert, and Maurizio Fava. 2009. Frontal EEG predictors of treatment outcome in major depressive disorder. *European Neuropsychopharmacology* 19, 11 (2009), 772–777.
[34] Ali Jalali, Pradeep Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. 2011. 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, 378–387.
[35] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2015. Multi-task, multi-kernel learning for estimating individual wellbeing. In *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, Vol. 898. 3.
[36] Houtan Jebelli, Byungjoo Choi, Hyeonseung Kim, and SangHyun Lee. 2018. Feasibility study of a wristband-type wearable sensor to understand construction workers' physical and mental status. In *Construction Research Congress*. 367–377.
[37] Maurice Jetté, Ken Sidney, and G Blümchen. 1990. Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clinical cardiology* 13, 8 (1990), 555–565.
[38] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. 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*. 7482–7491.
[39] Ronald C Kessler and Alex Luedtke. 2021. Pragmatic Precision Psychiatry—A New Direction for Optimizing Treatment Selection. *JAMA psychiatry* (2021).
[40] Jungyoon Kim, Jangwoon Park, and Jaehyun Park. 2020. Development of a statistical model to classify driving stress levels using galvanic skin responses. *Human Factors and Ergonomics in Manufacturing & Service Industries* 30, 5 (2020), 321–328.
[41] Meelim Kim, Jaeyeong Yang, Woo-Young Ahn, Hyung Jin Choi, et al. 2021. Machine Learning Analysis to Identify Digital Behavioral Phenotypes for Engagement and Health Outcome Efficacy of an mHealth Intervention for Obesity: Randomized Controlled Trial. *Journal of medical Internet research* 23, 6 (2021), e27218.
[42] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[43] Kent C Kowalski, Peter RE Crocker, and Nanette P Kowalski. 1997. Convergent validity of the physical activity questionnaire for adolescents. *Pediatric exercise science* 9, 4 (1997), 342–352.
[44] Richard L Kravitz, Naihua Duan, Sunita Vohra, Jiang Li, et al. 2014. Introduction to N-of-1 trials: indications and barriers. *Design and Implementation of N-of-1 Trials: A User's Guide* (2014), 1–11.
[45] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure.
[46] Michael Lecocke and Kenneth Hess. 2006. An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. *Cancer Informatics* 2 (2006), 117693510600200016.
[47] Dingwen Li, Jay Vaidya, Michael Wang, Ben Bush, Chenyang Lu, Marin Kollef, and Thomas Bailey. 2020. Feasibility Study of Monitoring Deterioration of Outpatients Using Multimodal Data Collected by Wearables. *ACM Transactions on Computing for Healthcare* 1, 1 (2020), 1–22.
[48] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. 2011. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized medicine* 8, 2 (2011), 161–173.
[49] Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. 2018. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–21.

[50] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.
[51] Nan Lv, Olusola A Ajilore, Corina R Ronneberg, Elizabeth M Venditti, Mark B Snowden, Philip W Lavori, Lan Xiao, Andrea N Goldstein-Piekarski, Joseph Wielgosz, Nancy E Wittels, et al. 2020. 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). *Contemporary clinical trials* 95 (2020), 106072.
[52] Jun Ma, Nan Lv, Lan Xiao, Andrea Goldstein-Piekarski, Joseph Wielgosz, Philip Lavori, Patrick Stetz, Lisa Goldman Rosas, Elizabeth Venditti, Mark Snowden, et al. 2020. 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.
[53] RD Mirza, S Punja, S Vohra, and G Guyatt. 2017. The history and development of N-of-1 trials. *Journal of the Royal Society of Medicine* 110, 8 (2017), 330–340.
[54] Kenneth E Mobily, Linda M Rubenstein, Jon H Lemke, Michael W O'Hara, and Robert B Wallace. 1996. Walking and depression in a cohort of older adults: The Iowa 65+ Rural Health Study. *Journal of Aging and Physical Activity* 4, 2 (1996), 119–135.
[55] Susan A Murphy and Derek Bingham. 2009. Screening experiments for developing dynamic treatment regimes. *J. Amer. Statist. Assoc.* 104, 485 (2009), 391–408.
[56] Ada Ng, Madhu Reddy, Alyson K Zalta, Stephen M Schueller, et al. 2018. Veterans' perspectives on fitbit use in treatment for post-traumatic stress disorder: an interview study. *JMIR mental health* 5, 2 (2018), e10415.
[57] T Christian North, PENNY McCullagh, Zung Vu Tran, David Ed Lavallee, Jean M Williams, Marc V Jones, and Anthony Col Papathomas. 2008. Effect of exercise on depression. (2008).
[58] World Health Organization et al. 2017. *Depression and other common mental disorders: global health estimates*. Technical Report. World Health Organization.
[59] Junbiao Pang, Qingming Huang, and Shuqiang Jiang. 2008. Multiple instance boost using graph embedding based decision stump for pedestrian detection. In *European conference on computer vision*. Springer, 541–552.
[60] David Paper and David Paper. 2020. Scikit-Learn Classifier Tuning from Simple Training Sets. *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python* (2020), 137–163.
[61] Meenal J Patel, Alexander Khalaf, and Howard J Aizenstein. 2016. Studying depression using imaging and machine learning methods. *NeuroImage: Clinical* 10 (2016), 115–123.
[62] Alfredo Raglio, Marcello Imbriani, Chiara Imbriani, Paola Baiardi, Sara Manzoni, Marta Gianotti, Mauro Castelli, Leonardo Vanneschi, Francisco Vico, and Luca Manzoni. 2020. Machine learning techniques to predict the effectiveness of music therapy: A randomized controlled trial. *Computer methods and programs in biomedicine* 185 (2020), 105160.
[63] Piyush Rai, Abhishek Kumar, and Hal Daume. 2012. Simultaneously leveraging output and task structures for multiple-output regression. *Advances in Neural Information Processing Systems* 25 (2012), 3185–3193.
[64] Amanda L Rebar, Robert Stanton, David Geard, Camille Short, Mitch J Duncan, and Corneel Vandelanotte. 2015. A meta-meta-analysis of the effect of physical activity on depression and anxiety in non-clinical adult populations. *Health psychology review* 9, 3 (2015), 366–378.
[65] Douglas K Russell. 1996. The Boltzmann distribution. *Journal of Chemical Education* 73, 4 (1996), 299.
[66] Yvan Saeys, Inaki Inza, and Pedro Larranaga. 2007. A review of feature selection techniques in bioinformatics. *bioinformatics* 23, 19 (2007), 2507–2517.
[67] Robert J Schalkoff. 2007. Pattern recognition. *Wiley Encyclopedia of Computer Science and Engineering* (2007).
[68] Fernando Seoane, Inmaculada Mohino-Herranz, Javier Ferreira, Lorena Alvarez, Ruben Buendia, David Ayllón, Cosme Llerena, and Roberto Gil-Pita. 2014. Wearable biomedical measurement systems for assessment of mental stress of combatants in real time. *Sensors* 14, 4 (2014), 7120–7141.
[69] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*. PMLR, 3145–3153.
[70] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine* 166, 10 (2006), 1092–1097.
[71] Andreas Ströhle. 2009. Physical activity, exercise, depression and anxiety disorders. *Journal of neural transmission* 116, 6 (2009), 777–784.
[72] John WG Tiller. 2013. Depression and anxiety. *The Medical Journal of Australia* 199, 6 (2013), S28–S31.
[73] Diane M Turner-Bowker, Martha S Bayliss, John E Ware, and Mark Kosinski. 2003. Usefulness of the SF-8™ Health Survey for comparing the impact of migraine and other conditions. *Quality of Life Research* 12, 8 (2003), 1003–1012.
[74] John Wallert, Emelie Gustafson, Claes Held, Guy Madison, Fredrika Norlund, Louise von Essen, and Erik Martin Gustaf Olsson. 2018. 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. *Journal of medical Internet research* 20, 10 (2018), e10754.
[75] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. 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*. 3–14.

[76] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[77] Leanne M Williams, Nicholas J Cooper, Stephen R Wisniewski, Justine M Gatt, Stephen H Koslow, Jayashri Kulkarni, Savannah DeVarney, Evian Gordon, and Augustus John Rush. 2012. Sensitivity, specificity, and predictive power of the "Brief Risk-resilience Index for Screening," a brief pan-diagnostic web screen for emotional health. *Brain and behavior* 2, 5 (2012), 576–589.

[78] Lan Yu, Daniel J Buysse, Anne Germain, Douglas E Moul, Angela Stover, Nathan E Dodds, Kelly L Johnston, and Paul A Pilkonis. 2012. Development of short forms from the PROMIS™ sleep disturbance and sleep-related impairment item banks. *Behavioral sleep medicine* 10, 1 (2012), 6–24.

[79] Yuezhou Zhang, Amos A Folarin, Shaoxiong Sun, Nicholas Cummins, Rebecca Bendayan, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, Petroula Laiou, et al. 2021. Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study. *JMIR mHealth and uHealth* 9, 4 (2021), e24604.