GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling

XUHAI XU, XIN LIU, and HAN ZHANG, University of Washington, USA
WEICHEN WANG and SUBIGYA NEPAL, Dartmouth College, USA
KEVIN S. KUEHN, University of Washington, USA
JEREMY F. HUCKINS, Dartmouth College, USA
MARGARET E. MORRIS and PAULA S. NURIUS, University of Washington, USA
EVE A. RISKIN, Stevens Institute of Technology, USA
SHWETAK PATEL and TIM ALTHOFF, University of Washington, USA
ANDREW CAMPBELL, Dartmouth College, USA
ANIND K. DEY and JENNIFER MANKOFF, University of Washington, USA

There is a growing body of research revealing that longitudinal passive sensing data from smartphones and wearable devices can capture daily behavior signals for human behavior modeling, such as depression detection. Most prior studies build and evaluate machine learning models using data collected from a single population. However, to ensure that a behavior model can work for a larger group of users, its generalizability needs to be verified on multiple datasets from different populations. We present the first work evaluating cross-dataset generalizability of longitudinal behavior models, using depression detection as an application. We collect multiple longitudinal passive mobile sensing datasets with over 500 users from two institutes over a two-year span, leading to four institute-year datasets. Using the datasets, we closely re-implement and evaluated nine prior depression detection algorithms. Our experiment reveals the lack of model generalizability of these methods. We also implement eight recently popular domain generalization algorithms from the machine learning community. Our results indicate that these methods also do not generalize well on our datasets, with barely any advantage over the naive baseline of guessing the majority. We then present two new algorithms with better generalizability. Our new algorithm, *Reorder*, significantly and consistently outperforms existing methods on most cross-dataset generalization setups. However, the overall advantage is incremental and still has great room for improvement. Our analysis reveals that the individual differences (both within and between populations) may play the most important role in the cross-dataset generalization challenge. Finally, we provide an open-source benchmark platform **GLOBEM** – short for **G**eneralization of **LO**ngitudinal **BE**havior **M**odeling – to consolidate all 19 algorithms. GLOBEM can support researchers in using, developing, and evaluating different longitudinal behavior modeling methods. We call for researchers' attention to model generalizability evaluation for future longitudinal human behavior modeling studies.

GLOBEM website: the-globem.github.io | GLOBEM codebase: github.com/UW-EXP/GLOBEM

Authors' addresses: Xuhai Xu, xuhaixu@uw.edu; Xin Liu; Han Zhang, University of Washington, Seattle, WA, USA; Weichen Wang; Subigya Nepal, Dartmouth College, Hanover, NH, USA; Kevin S. Kuehn, University of Washington, USA; Jeremy F. Huckins, Dartmouth College, USA; Margaret E. Morris; Paula S. Nurius, University of Washington, USA; Eve A. Riskin, Stevens Institute of Technology, USA; Shwetak Patel; Tim Althoff, University of Washington, USA; Andrew Campbell, Dartmouth College, USA; Anind K. Dey; Jennifer Mankoff, University of Washington, USA.

2474-9567/2022/12-ART190
https://doi.org/10.1145/3569485

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Applied computing → Life and medical sciences**.

Additional Key Words and Phrases: Generalizability, Behavior Modeling, Passive Sensing

ACM Reference Format:
Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2022. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 190 (December 2022), 32 pages. https://doi.org/10.1145/3569485

1 INTRODUCTION

Ubiquitous computing is entering almost every aspect of our life. As close companions to users, smartphones and wearable devices can continuously and passively capture various aspects of daily behavior. In the past decade, a large body of work has demonstrated the capability of longitudinal passive sensing and behavior modeling in many application areas, such as detecting physical health issues [5, 60, 65, 104], monitoring mental health status [96, 101], measuring job performance [57, 61], tracking education outcomes [97, 114], and tracing social justice [82]. Researchers have used a variety of approaches to address their research questions, ranging from doing statistical analysis to building machine learning (ML) models.

Due to the high cost of conducting longitudinal passive sensing studies, most prior research collected self-reported surveys data (as ground truth) and sensor data from one population over a few weeks or months. Their data analysis and models mainly focused on a single dataset. However, to build a model with practical and useful deployability, it is essential for researchers to evaluate the model across multiple datasets to ensure its generalizability. As real-life deployment will introduce new populations under new contexts, the model needs to generalize well to unseen data with robust performance. The gap between the single dataset-based analysis and the need for cross-dataset evaluation has been long-standing in the longitudinal behavior modeling domain. Moreover, the generalization challenge could come from multiple sources, such as different populations, different users in the same population, and the same users across different time periods. It remains unknown to the community which of these issues makes cross-dataset generalization particularly challenging. In this work, we take the first step towards a systematic evaluation of behavior models' generalizability[1] across multiple datasets, using depression detection as an example. We call for researchers' attention to incorporate this critical step into model analysis pipeline in the future.

To perform an analysis of model generalizability, we formed a collaboration of two research groups across two institutes. Each group conducted two rounds of longitudinal passive sensing studies (one in 2018 and one in 2019, before the impact of the COVID-19 pandemic), focusing on depressive symptom detection. At each institute, there was a small fraction of overlapping users across the two years. We employed a uniform data transformation and feature extraction process to build four institute-year datasets, and re-implemented a large proportion of behavior features in prior depression detection studies (*e.g.*, [19, 80, 93, 98, 100, 101]). We then evaluated the cross-dataset generalizability of behavior models using the four datasets. In this paper, we focused on a common binary classification task as a starting point: distinguishing whether participants had at least mild depression.

We closely re-implemented 9 prior depression detection algorithms using overlapping features from the four datasets, and tested the models generated via these algorithms, using a consistent cross-validation setup on each dataset. Although these models outperform the naive baseline predicting the majority ($\Delta = 7.0 \pm 7.6\%\%$), we

[1]Our setup of the domain generalization tasks on passive sensing data aligns with ML community's definition of *generalizability* as an ML model's generalization ability, that is, how well a model can adapt properly to new, previously unseen data [63, 68]. This term is also closely related to one type of *reproducibility* of scientific research: conceptual reproducibility, the ability to obtain similar findings under new conditions that match the theoretical description of the original experiment [37, 58] In this paper, we follow a few recent ML publications [1, 110] and use the term *generalizability*.

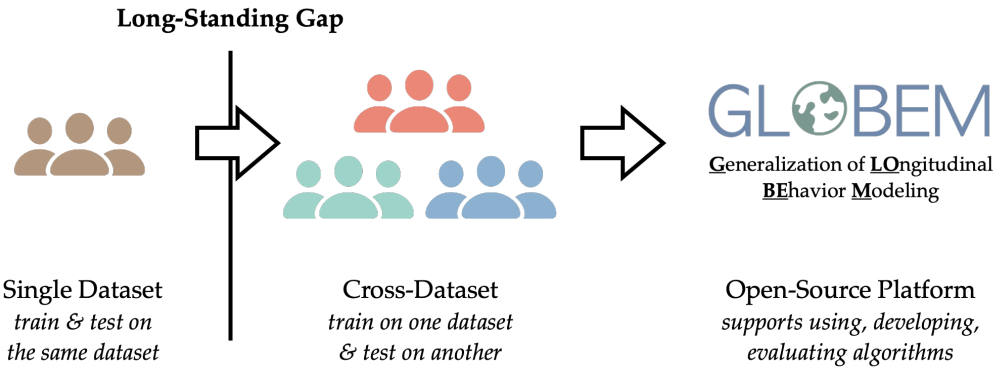


Fig. 1. Overview of The Contributions of This Work. We systematically evaluate cross-dataset generalizability of 19 algorithms: 9 prior behavior modeling algorithm for depression detection, 8 recent domain generalization algorithms, and 2 two new algorithms proposed in this paper. Our open-source platform GLOBEM consolidates these 19 algorithms and support using, developing, evaluating various algorithms.

observed a substantial drop between these models' performance on our datasets and their reported results in literature ($\Delta = 18.9 \pm 9.7\%$), which indicates the lack of cross-dataset generalizability of these algorithms.

Second, a range of domain generalization algorithms were developed with recent advances in the ML community [94, 118], which were mainly designed for computer vision (CV) or natural language processing (NLP) tasks. However, there is no prior work applying these methods to the longitudinal behavior modeling domain. To test these methods, we re-implemented 8 types of recent deep-learning-based domain generalization algorithms, and built 15 different models using these algorithms (an algorithm can generate multiple models). We used a leave-one-dataset-out setup to evaluate the cross-dataset generalizability of prior depression detection methods and recent domain generalization algorithms, *i.e.*, using one dataset as the testing set and the rest of datasets as the training set. However, our experimental results revealed that these models did not generalize well to our longitudinal passive sensing domain. The best model only has a minimal advantage on ROC AUC (54.1%) over the naive majority baseline (50.0%, $\Delta = 4.1\%$). It is worth noting that our main goal is not to criticize or single out any prior work (including our own), but to emphasize the importance and lack of model generalizability for longitudinal passive sensing of human behaviors.

We further present two new algorithms to enhance model generalizability. The first method, *Clustering*, extends a Siamese network and builds an ensemble model based on unsupervised clustering, which could utilize the similarity of behavior trajectories among individuals. The second method, *Reorder*, creates a new task of solving a temporal reordering puzzle in addition to the main depression detection task, which was inspired by self-supervised techniques in CV and NLP approaches [24, 77]. *Reorder* is forced to learn the continuity of behavior trajectories and achieve better generalization. Our results demonstrate that *Reorder* significantly outperforms other methods by at least 3.4% on ROC AUC (6.3% relative advantage) and 3.2% on balanced accuracy (6.2% relative advantage) in the leave-one-dataset-out evaluation. Moreover, our datasets support multiple cross-dataset tasks besides leave-one-dataset-out, including different populations (two institutes), different users within the same population (same institute across two years), and same users at different times (overlapping users across two years). *Reorder* consistently achieves the best or the second best performance under other cross-institute and cross-year generalization tasks. Meanwhile, *Clustering* achieves the highest upper bound when the model training can early stop at the optimal epoch [112]. These results indicate the strong generalizability advantage of our proposed techniques. Nevertheless, we acknowledge that the advantage of our new methods is statistically significant yet incremental. More future work will be needed to further advance the area.

In addition, the comparison between multiple generalization tasks enables us to investigate the major challenge of domain generalization in our dataset. We identify that compared to the time period difference (for the same user), individual difference (*i.e.*, user-user difference, either from the same or different populations) is the main source of various generalization challenges.

We consolidated our implementation of 19 methods (9 prior depression detection algorithms + 8 recent domain generalization algorithms + 2 new algorithms) and developed a benchmark platform, **GLOBEM** (short for **G**eneralization of **LO**ngitudinal **BE**havior **M**odeling), to accelerate cross-dataset evaluation by others in future research. Compared to existing cross-dataset generalization platforms (DomainBed [36], DeepDG [94]) and benchmarks (WILDS [48]), our platform further extends the domain generalization problem to the longitudinal passive sensing field. GLOBEM not only supports flexible and rapid evaluation of the existing methods, but also provides easy-to-extend templates for researchers to develop and prototype their own algorithms and systems.

The contributions of our work can be summarized as follows:

- We highlight the importance of cross-dataset generalizability evaluation of longitudinal behavior models. Joining forces across two research groups, we collected four passive sensing datasets with similar sensors and labels, using depression detection as an example. To the best of our knowledge, we are the first to conduct a comprehensive multi-dataset evaluation of existing behavior modeling algorithms.
- We re-implemented 9 prior depression detection algorithms and 8 modern domain generalization algorithms. Our experiments revealed the general lack of generalizability of these methods in our data format.
- We present two new algorithms, *Clustering* and *Reorder*. Our results demonstrated the advantage of *Reorder* under various cross-dataset setups, as well as the potential of *Clustering* under the assumption of knowing the optimal training epoch.
- Our multiple cross-dataset evaluation setups (cross institutes, cross years) reveal that individual behavior difference is the major source of model generalization challenge in our area.
- We built and open-sourced the first cross-dataset benchmark platform, GLOBEM, in the longitudinal behavior modeling domain. It incorporated all 19 algorithms (9+8+2) to support future researchers in testing existing methods and implementing new algorithms.

We envision our work can inspire researchers in the ubiquitous computing and ML communities to put continued and additional effort into building behavior modeling algorithms that are robust and generalizable across multiple domains and datasets.

2 BACKGROUND

In this section, we first briefly summarize prior work on depression detection based on longitudinal passive sensing (Section 2.1). We then provide a high-level overview of the recent advances in the ML community on domain generalization (Section 2.2).

2.1 Depression and Passive Sensing

Depression, formally known as major depressive disorder (MDD), is a common and important mental health problem that affects people worldwide. Research has found that depression affects approximately 216 million people globally [92]. A recent report estimated that 7.2% of all U.S. adults had at least one depressive episode in a year, *i.e.*, a period of more than two weeks of depressive symptoms [72], and the prevalence has increased during the pandemic [44, 66]. Early detection of depression can potentially assist in mitigating or preventing its detrimental consequences. There is a growing realization that everyday devices can help us understand the relationship between people's daily actions and depressive symptoms, by continuously and passively collecting

daily behavioral data [4, 10, 14]. The success in the last decade of using mobile sensing for depression research has attracted increasing attention of researchers from various communities [49, 103].

Earlier research focused on understanding the statistical relationship, such as correlation, between depressive symptoms and mobile sensing data [11, 80]. For example, Saeb et al. [80] identified a significant correlation between depression scores and features related to location and phone usage. Ben-Zeev et al. [11] found a significant correlation between changes in depression severity levels and features related to sleep duration, speech duration, and mobility. Recently, researchers have leveraged the results of correlation analysis to build ML models for depression diagnosis and detection [19, 69, 93]. For instance, Farhan et al. [27] employed location features to detect biweekly depression and their top model achieved an F1 score of 82.0% on an eight-month dataset with 79 college students. Wang et al. [98] hand-crafted several cross-sensor features using data from mobile and wearable devices. On a dataset collected from 68 college students across two nine-week terms, their best model achieved 81.5% recall and 69.1% precision. Xu et al. [100] proposed an automated feature extraction pipeline across multiple sensors' data. They used association rule mining to discover frequent behavior rules and extract interpretable features. Their model achieved an accuracy of 81.8% and an F1 score of 84.3% on a dataset with 138 college students over 16 weeks. They further developed a personalized depression detection model using a new method based on collaborative-filtering [101]. Their model achieved an accuracy of 82.4% and an F1 score of 85.5% on a similar dataset.

However, most existing depression detection research have mainly developed and evaluated their models on only a single dataset. There is a lack of understanding and evaluation of these models' generalizability across multiple datasets. We aim to address this gap in this paper.

2.2 Cross-dataset Generalization

Building a model that can generalize across multiple domains or datasets has been a challenging problem in the ML community. This is one of the problems that transfer learning aims to address, namely domain adaption, which focuses on cases where the model can access some data of the target domains (*e.g.*, datasets) in addition to the data of the source domains [74, 95]. A more challenging task is domain generalization, where the model can only access source domains [118]. In this paper, we focus on the domain generalization task with datasets that have homogeneous feature spaces and label spaces but with divergence [115]. Such a task corresponds to an ideal real-life deployment setup: directly applying a trained model to new users, without the need to access any data from the new users.

In the past few years, ML researchers have proposed a wide range of algorithms for domain generalization. Most of them belong to one of the following three categories [94]: 1) Data manipulation, which enhances the data by augmentation or generation techniques to assist the model training (*e.g.*, [23, 113]); 2) Representation learning, which aims to learn desirable feature representations that can generalize across domains (*e.g.*, [2, 28, 31]); 3) Learning strategy, which focuses on exploiting the training procedure to promote a model's generalizability (*e.g.*, [50, 88, 111]). We refer readers to recent domain generalization surveys for more details [94, 118]. Almost any applied ML area will encounter the cross-dataset domain generalization challenge for real-life deployment, such as object recognition [6, 15, 45, 86], NLP [70, 117], affective computing [20, 42], security [41, 116], and intelligent interaction [102, 105, 106]. Researchers have developed cross-dataset benchmark platforms such as DomainBed [36], DeepDG [94], and WILDS [48] to facilitate related studies in the ML community. Recently, researchers started to merge multiple passive sensing datasets for behavior model training [1]. However, there is no prior work on cross-dataset generalizability evaluation in the longitudinal passive mobile sensing area. Perhaps the most relevant work is by Mishra *et al.* [62]. They evaluated three models' generalizability in the physiological stress detection domain using four datasets collected from different wearable devices in a passive manner. However, their models aimed at short-term detection (over a few seconds), so the ground truth labels

were frequent and rich. Moreover, the datasets they employed were all collected in a lab setting. Instead, we focused on in-the-wild longitudinal passive sensing problems, where the ground truth is usually much more sparse (*e.g.*, once per one to two weeks or less frequent for depression labels [14, 100]), and the field experiments were not controlled. To the best of our knowledge, we are the first to conduct an extensive cross-dataset evaluation analysis in our area.

3 DATASET

We combined passive sensing data from two institutes collected by two research groups, each across two years (before the impact of the pandemic). We first introduce the data collection process (Section 3.1), and then describe the feature extraction steps that were employed consistently across all of the datasets (Section 3.2).

3.1 Data Collection

Both research groups employed the data collection model proposed in [96]. The datasets were collected from two different Carnegie-classified R-1 universities in the United States using a similar method [43, 82]. In both institutes, the data collections went through an IRB review and approval. We recruited undergraduate students via emails and social media posts. We collected four datasets in total, two from each institute (one in 2018, and one in 2019). Every dataset contained data from participants over the same academic quarter for 10 weeks (Spring quarter, approximately from late-March to mid-June) [2], so the impact of seasonal effects was controlled. We summarize the study details in Table 1.

The four datasets have 155, 218, 93, and 152 students (618 person-years in total), respectively, with a high percentage of female students. There were a few overlapping participants who were in both years' datasets (23 in Institute 1, and 58 in Institute 2); thus the datasets contain 534 unique participants in total. Every quarter,

Table 1. Basic Study Information and Participant Demographics of Four Datasets. Participants with less than 2 weekly EMAs or less than a quarter of their sensor data (*i.e.*, missing rate > 75%) were excluded from the dataset. In the ground truth row, the percentage in the parentheses indicates the proportion of participants having at least mild depressive symptoms based on the corresponding questionnaires. Gender acronym - F: Female, M: Male, NB: Non-binary. Racial acronym - A: Asian, B: Black or African American, H: Hispanic, N: American Indian/Alaska Native, PI: Pacific Islander, W: White, NA: Did not report. & is used when participants reported more than one races.

	Institute1		**Institute2**	
	Year1 - DS1	**Year2 - DS2**	**Year1 - DS3**	**Year2 - DS4**
Participants	- Total: 155 - Gender: F 107, M 48 - Race: A 82, B 5, H 9, N 4, PI 3, W 50, A&PI 2	- Total: 218 - Gender: F 111, M 107 - Race: A 102, B 6, H 10, N 2, PI 1, W 70, A&B 1, A&W 16, H&W 2, B&W 2, A&H&W 1, B&H&W 1, H&N&W 1, NA 3 - 23 also in Year1	- Total: 93 - Gender: F 65, M 27, NB 1 - Race: A 20, B 3, H 2, N 4, PI 1, W 52, B&H 1, H&W 5, NA 5	- Total: 152 - Gender: F 101, M 49, NB 2 - Race: A 28, B 2, H 4, N 4, W 100, B&H 1, NA 13 - 58 also in Year1
Ground Truth	- Weekly: Depression & Affect (44.4%) - End-term: BDI-II (35.4%)	- Weekly: PHQ-4 (50.3%) - End-term: BDI-II (42.9%)	- Weekly: PHQ-4 (35.9%) - End-term: PHQ-4 (42.4%)	- Weekly: PHQ-4 (37.7%) - End-term: PHQ-4 (33.8%)
Sensor Data	- Overlap: Location, Phone Usage - Compatible: Physical Activity (Fitbit), Sleep (Fitbit) - Incompatible: Bluetooth, WiFi, Call, Battery		- Overlap: Location, Phone Usage - Compatible: Physical Activity (phone), Sleep (phone) - Incompatible: Audio	

[2]Our Institute1Year1 dataset contains two quarters' data. To make it consistent with other datasets, we took the Spring quarter for our analysis.

participants received up to $245 (Institute 1) and $100 (Institute 2) for compensation, based on their compliance. The studies at Institute 1 also used a wearable Fitbit for data collection, and participants were allowed to keep the Fitbit after the study if their study compliance was high enough.

3.1.1 Ground Truth. In both institutes, we employed well-established and validated questionnaires to obtain ground truth of participants' depressive status. We collected survey results on a weekly basis, and an end-term survey at the end of the quarter. These surveys were delivered via smartphones. The weekly surveys include PANAS [22] (Institute1Year1 only), PHQ-4 [54] (remaining datasets). The end-term surveys include BDI-II [9] (Institute1) and PHQ-4 (Institute2).

PHQ-4 and BDI-II are both screening tools for further inquiry of clinical depression diagnosis. They distinguish four severity levels: no or least, mild, moderate, and severe depressive symptoms. As an initial step of model generalizability evaluation, we focus on a binary classification task to distinguish whether participants' scores indicate at least mild depressive symptoms through the scales (*i.e.*, PHQ-4 > 2, BDI-II > 13)[3].

Note that although PANAS contains questions related to depressive symptoms (*e.g.*, "distressed"), it does not have a comparable theoretical foundation for depression detection like PHQ-4 or BDI-II. Therefore, to maximize the compatibility of the datasets, we trained a small decision tree (depth=2) on Institute1Year2 data that has both PANAS and PHQ-4 scores to generate reliable ground truth labels. We used PNANS scores (on a 1-5 Likert scale) on two affect questions, depressed and nervous, as the input and PHQ-4 score-based depression binary label as the output. Our model achieved an accuracy of 74.5% and an F1-score of 76.3% on a 5-fold cross-validation. It generated a simple rule: the user would be labeled as having at least mild depression when the distress score is greater than 1, or the nervous score is greater than 2. We then applied this rule to Institute1Year1 dataset to generate labels. We discuss potential limitations of our label generation method in Section 6. After this procedure, all datasets contain binary depression labels for both end-term and repeated weekly prediction tasks. The average number of total depression labels is 12.9 ± 4.1 per person.

3.1.2 Passive Sensing Data. The two research groups employed two separate smartphone applications for passive sensing data collection. In Institute 1, the app was developed based on the AWARE Framework [29]. In Institute 2, the app was developed based on the work by Wang et al. [96]. Both apps are compatible with iOS and Android systems. The functionalities of these two apps are very similar: they both could collect location, phone usage (screen status), physical activities, and sleep patterns. The app of Institute 1 also recorded a phone's nearby Bluetooth and WiFi addresses, as well as call logs. The app of Institute 2 further recorded audio data via microphone. Moreover, the studies in Institute 1 employed a wearable Fitbit to collect participants' detailed activities and sleep behaviors, while the studies in Institute 2 combined multiple sensors (*e.g.*, accelerometer, gyroscope, microphone, and lightness sensor) to infer activity and sleep behaviors [18, 96].

3.2 Feature Extraction

To ensure the consistency and reproducibility of features extracted from the multiple datasets, we utilized the RAPIDS [91] platform that supports feature extraction from multiple passive mobile sensors with flexible time windows. RAPIDS is open source to enable Reproducible Analysis Pipeline for Data Streams. It supports data streams logged by multiple platforms (smartphones, Fitbit wearables, and Empatica wearables), integrates multiple previous research efforts (*e.g.*, [8, 14, 25]), and has been leveraged in a range of passive sensing and behavior modeling studies (*e.g.*, [90, 109]). As RAPIDS builds a feature extraction pipeline based on data collected by the AWARE framework, we transformed Institute 2's data format to make it compatible with RAPIDS, without altering the raw data content. We used the same RAPIDS feature extraction configuration across all datasets.

[3]For simplicity, we use "with/without depression" throughout the paper. It is worth noting that this does not imply a clinical diagnosis.

One of our goals is to reproduce a list of previous depression detection models. Therefore, our feature extraction focuses on re-implementing features that occur in previous work. RAPIDS already provides a wide range of features that we can easily leverage[4]. We extended it to cover features of the four most important modalities across the four datasets: location, phone usage, physical activity, and sleep behaviors. We excluded features that were not common in all datasets to ensure maximum compatibility, including features related to Bluetooth, WiFi and calls in DS1 and DS2, as well as audio features in DS3 and DS4.

3.2.1 Data Type: Location. We included the all location features provided by RAPIDS, including travel distance, location variance, location entropy, moving speed, number of frequent locations visited, radius of gyration, *etc.* We also incorporated more location features (duration of staying and percentage of total time) related to specific points of interest, including places for living/home, study, exercise, and relaxation.

3.2.2 Data Type: Phone Usage. In addition to screen-related features in RAPIDS about the statistics of unlocking episodes (count, sum, mean, standard deviation, maximum, minimum), we further considered these features at different locations (home and study places) to contextualize the phone usage behavior.

3.2.3 Data Type: Physical Activity. We leveraged the Fitbit-based activity features in RAPIDS, which include 1) high-level summary features like the number of steps, duration of being active, 2) low-level features that consider the statistics (mean, standard deviation, maximum, minimum) of active or sedentary episodes.

3.2.4 Data Type: Sleep. We utilized the Fitbit-based sleep features in RAPIDS. Similar to activity features, sleep features also include 1) high-level summary features such as total duration of being asleep or in bed, 2) low-level features that calculate the statistics (count, mean, maximum, minimum) of episodes of being asleep, restless, and awake during the night.

3.2.5 Data Time Range. Different prior studies extract features at different frequencies (from an epoch of a day to multiple days). We incorporated all of them when extracting the features in Section 3.2.1-3.2.4, including four epochs of a day (morning 6 am - 12 pm, afternoon 12 pm - 6 pm, evening 6 pm - 12 am, and night 12 am - 6 am) [19, 82, 101], the whole day [19, 80, 93, 98], and past two weeks [14, 27, 55, 93].

3.2.6 Post-processing. After a consistent feature extraction process, we prepared four datasets with the same feature and label spaces. Some prior work also involved the procedure of feature normalization, so we doubled the feature number by adding normalized features based on each individual's distribution: subtracting the median and scaling with the 5-95 quantile range on each individual. The current data format for each participant is a time-series feature-vector data (*i.e.*, a long feature matrix), accompanied by a short list of labels on certain dates. Since participants may have a different number of labels, we further sliced the feature sequence based on labels to make consistent data input shapes. Specifically, given a label collected on a particular day, we sliced a feature matrix of the past four weeks, counting back from the day (28 days × feature number). We picked four weeks in order to cover previous depression detection models' feature calculations [14, 27, 55]. After the slicing, every participant label corresponded to an input feature matrix with the same shape. Note that missing features (due to the missing raw data, accounting for 7.3±2.3% of the data) were imputed with the median value. Once we finished post-processing the feature data, we could then evaluate various algorithms across these datasets[5].

[4]Please refer to https://www.rapids.science/1.6/features/feature-introduction/ for features in each data type.

[5]Due to IRB requirements, we are not able to release the full data analyzed in this article. However in [107], we release a data set that includes and extends the data from Institute 1 analyzed in this article.

4 METHODS AND BENCHMARK

We developed a uniform benchmark platform and compared 19 methods' generalizability on our four datasets. We first introduce prior behavior models that focused on depression detection in Section 4.1. We then describe the eight recent deep learning-based domain generalization algorithms in Section 4.2. We highlight our newly proposed methods that demonstrated improved generalizability in Section 4.3. All of these methods were evaluated on a uniform platform that we developed and plan to open-source (Section 4.4).

4.1 Existing Depression Detection Algorithms

We introduce the nine prior depression detection algorithms with technical details. We re-implemented the algorithms with the same or similar features, and evaluated these models on our datasets. [6] The specific model setup can be found in the configuration files in our codebase.

(1) **Canzian *et al.*** [14]: used location trajectory features directly computed from the past two-week time window to train a support vector machine (SVM) for depression detection.

(2) **Saeb *et al.*** [80]: used the combination of location and screen features and aggregated their daily average of the past two weeks to train a logistic regression model with elastic regularization.

(3) **Farhan *et al.*** [27]: used location and physical activity features from the past two-week window to train an SVM model.

(4) **Wahle *et al.*** [93]: used features from several sensors (activity, location, WiFi, screen, and call) over the past two weeks. They used both daily aggregation (*i.e.*, mean, sum, variance) and direct computation of the features of the two weeks to build SVM and Random Forest models. We left out the WiFi and call features to ensure its compatibility with our datasets.

(5) **Lu *et al.*** [55]: used location, activity, and sleep features computed from the past two weeks and built multi-task learning models combining linear regression and logistic regression. To further deal with device platform differences, they built one model for iOS devices and one for Android devices.

(6) **Wang *et al.*** [98]: used location, screen, activity, sleep, and audio features and aggregated their daily average and slope of the past two weeks (for the frequent prediction) or the whole study period (for the end-of-term prediction). They built a lasso-regularized logistic regression model for the prediction. We excluded audio features as they were not collected in all datasets.

(7) **Xu *et al.*- Interpretable** [100]: used location, screen, activity, and sleep features in multiple epochs of a day (morning, afternoon, evening, night). They first applied association rule mining to mine out interpretable behavior rules that capture differences between participants with depression and without depression. Then, they used the rules to filter and aggregate features of multiple days and built an Adaboost model for the detection.

(8) **Xu *et al.*- Personalized** [101]: used a similar set of features as [100]. With each feature as a time sequence, they computed a user behavior relevance matrix using the square of Pearson correlation to capture users with strong positive or negative correlation. They used a traditional collaborative-filtering-based model to select features and obtain an intermediate prediction using each feature, and combined the results of all features via majority voting.

(9) **Chikersal *et al.*** [19]: used a similar set of basic features as [100] and calculated more aggregations (breakpoint and slope) across multiple time ranges (daily and biweekly). They first trained a nested randomized logistic regression for feature selection. Then, they trained separate gradient boosting and

[6] It is worth noting that our re-implementation may not be exactly the same as the prior work due to the lack of open-source code.

logistic regression models using data from every sensor, and combined the prediction with another Adaboost model to generate the final prediction.

It is worth noting that *Xu et al.- Interpretable* was originally developed on the same raw data of DS1, and *Xu et al.- Personalized* was developed on DS1 and DS2, using a more strict data filtering criteria (resulting in a smaller user group). However, the original work employed additional feature types (Bluetooth, WiFi, call, message, and battery) and these features were not included in this work, to allow for comparisons across all of the datasets.

4.2 Recent Domain Generalization Algorithms

Previous depression detection algorithms did not focus on the domain generalization challenge. In the ML community [94, 118], a range of domain generalization algorithms were proposed recently, but they were mostly designed for CV/NLP tasks. There is no prior work evaluating these methods on longitudinal passive sensing datasets. Therefore, we implemented eight well-studied deep learning techniques to cover the major approaches of domain generalization [94], including 1) data manipulation (Mixup [113]), 2) representation learning (IRM [2], DANN [31], CSD [76]), and 3) learning strategy (MLDG [50], MASF [26], Siamese [47]). The input-output format of these models is the same: we picked a subset of important daily features in the most recent traditional depression detection algorithms [19, 101]. We then used the past-four-week feature matrix as the input, and the depression label as the output.

(1) **ERM** (Empirical Risk Minimization) [89]: the basic model training techniques without particular design for domain generalization. ERM shows a competitive performance in previous CV generalization tasks [36, 94]. We implemented multiple architectures with ERM: a) **ERM-1D-CNN**: one-dimensional CNN that treats the data as a time series of length 28; b) **ERM-2D-CNN**: two-dimensional CNN that treats the data axs an one-channel image; c) **ERM-LSTM**: another architecture to model time-series data; d) **ERM-Transformer**: a transformer-based architecture for modeling sequence data.

(2) **Mixup** (ERM-Mixup) [113]: a popular data manipulation and augmentation technique that performs linear interpolation between any two instances with a weight sampled from a Beta distribution. Mixup can be plugged into any model architecture and training pipeline. In this paper, we used 1D-CNN in our experiment as it has a similar performance as 2D-CNN (as shown in Section 5), while more robust to feature positions in the feature matrix. Similarly, we also used 1D-CNN for the other methods in the rest of this section if they are agnostic of architectures.

(3) **IRM** (Invariant Risk Minimization) [2]: a representation learning paradigm to estimate invariant correlations across multiple distributions and learn a data representation such that the optimal classifier can match all training distributions.

(4) **DANN** (Domain-Adversarial Neural Network) [31]: another representation learning technique that adversarially trains the generator and discriminator. The discriminator is trained to distinguish different domains, while the generator is trained to fool the discriminator to learn domain-invariant feature representations. For our purposes, we treated each dataset as a domain (**DANN - Dataset as Domain**), or each person as a domain (**DANN - Person as Domain**).

(5) **CSD** (Common Specific Decomposition) [76]: a feature disentanglement-based representation learning technique from the multi-component analysis perspective, which extracts the domain-shared and domain-specific features using separate network parameters. Similar to DANN, we also investigated two versions of domain: **CSD - Dataset as Domain**, and **CSD - Person as Domain**.

(6) **MLDG** (Meta-Learning for Domain Generalization) [50]: one of the first methods using meta-learning strategy for domain generalization. MLDG splits the data of the training domains into meta-train and

meta-test to simulate the domain shift to learn general features. We again tried **MLDG - Dataset as Domain**, and **MLDG - Person as Domain**.

(7) **MASF** (Model-Agnostic Learning of Semantic Features) [26]: a learning strategy that combines meta-learning and feature disentanglement. After simulating domain shift by domain split, MASF further regularizes the semantic structure of the feature space by introducing a global loss (to preserve relationships between classes) and a local loss (to promote domain-independent class clustering). We tested **MASF - Dataset as Domain**, and **MASF - Person as Domain**.

(8) **Siamese Network** [47]: a metric-learning based strategy to find a better pair-wise distance metric. It aims to decrease the distance between positive pairs (*i.e.*, same labels) and increase the distance between negative pairs (*i.e.*, different labels).

4.3 New Algorithms Leveraging Longitudinal Behavior Trajectory

We observe gaps in the existing algorithms. First, prior depression detection algorithms (Section 4.1) did not attempt to address the domain generalization challenge. In addition, recent domain generalization models proposed by the ML community (Section 4.2) were not specifically designed for longitudinal passive sensing data. Therefore, we proposed two new methods that leverage human behavior characteristics from two different perspectives: one using the similarity of behavior trajectories among different individuals (Section 4.3.1), and the other leveraging the continuity of behavior trajectories of each individual (Section 4.3.2).

4.3.1 Clustering Similar Behavior Trajectories to Train Independent Models. This method stems from a simple intuition: Although participants came from different populations, certain periods of some users' behavior trajectories could be similar and clustered together. Data in the same cluster might have closer distributions and simplify the classification task within the cluster.

Specifically, we built an ensemble model based on unsupervised clustering. Using all training data, we first followed [38, 99] to train a deep clustering model with convolutional auto-encoders (DCEC) that assigns data points with cluster indices. For each cluster, we then used data in that cluster to train a small Siamese model [47]. In the inference stage, data would be fed into the DCEC model to find the cluster index, and classified by the corresponding Siamese model.

4.3.2 Reordering Behavior Trajectory to Force The Model to Learn Behavior Continuity. The challenge of domain generalization is mainly caused by the data distribution shift in heterogeneous domains. In our case, such a shift comes not only from dataset differences (*i.e.*, each subpopulation behavior pattern varies [21, 59]), but also from individual differences (*i.e.*, each person behaves uniquely [67, 83]). However, despite these differences, there still exists a range of similarities among individuals' behaviors. For example, people tend to have daily routines, which define the structure of and influence of almost every aspect of everyday behaviors [7]. Although individuals have unique routines, these patterns would lead to *continuous* or even repetitive behavior trajectories from day to day [40]. Such an observation motivates us to leverage the behavior continuity and construct a self-supervised learning task to obtain generalizable feature representations.

Self-supervised learning is a recently popular learning paradigm that builds self-supervised tasks (*i.e.*, pretext tasks) from the unlabeled data. The pretext tasks are often not directly related to the main prediction task. It has been applied to domain generalization problems in CV tasks (*e.g.*, JiGen [16], SelfReg [46]) and NLP tasks (*e.g.*, BERT [24]). For example, BERT used a pretext task where words in a text are randomly masked and the goal is to predict these words.

To leverage the continuity of behavior trajectory, inspired by [16], we proposed a new multi-task learning model, **Reorder**, with a new pretext task called reordering puzzle (see Figure 2): we shuffled the temporal order of the feature matrix, and trained a model to reconstruct the original sequence, jointly optimized with the main

classification task over different domains. The model needs to achieve two tasks simultaneously: 1) it will learn to capture the continuity of behavior trajectories, so that it can find the right temporal order of the time-series feature data after shuffling; and 2) it will also learn to solve the main task (*i.e.*, depression detection in our case). Due to the prevalence of the continuous behavior trajectories based on human nature (analogous to the continuous edges and patterns in images [16]), solving the first task by learning such continuity could assist the model to extract more generalizable representations of behavior trajectories across individuals. This could potentially enable more robust domain generalization in behavior modeling.

Specifically, we created a multi-task learning model function h, with the 1D-CNN based embedding (parameters θ_f), fully connected layers for reordering (parameters θ_r), and fully connected layers for classification (parameters θ_c). The first task is the main classification task. The loss function of this task is $\mathcal{L}_c(h(x|\theta_f, \theta_c), y)$, where x is the input matrix, and y is the classification label. The second task is the reordering task. We first sliced the feature matrix along the temporal dimension into n segments and then shuffled these segments. We picked the number of segments $n = 10$ ($\lceil 28/3 \rceil$) since 28! or 14! (28/2) is too computationally expensive. Moreover, as 10! total possible permutations is an overly large number, we followed the practice in [16] and predetermined a subset of $P = 200$ permutations by following the Hamming-distance-based method [71]. We then assigned an index to each permutation. Within the subset, the reordering task is equivalent to identifying the index of the permutation, which is essentially a classification task. Therefore, the loss function of the reordering task is $\mathcal{L}_r(h(z|\theta_f, \theta_r), p)$, where z is the feature matrix x after the reordering, and p is the permutation index. Overall, the model can be

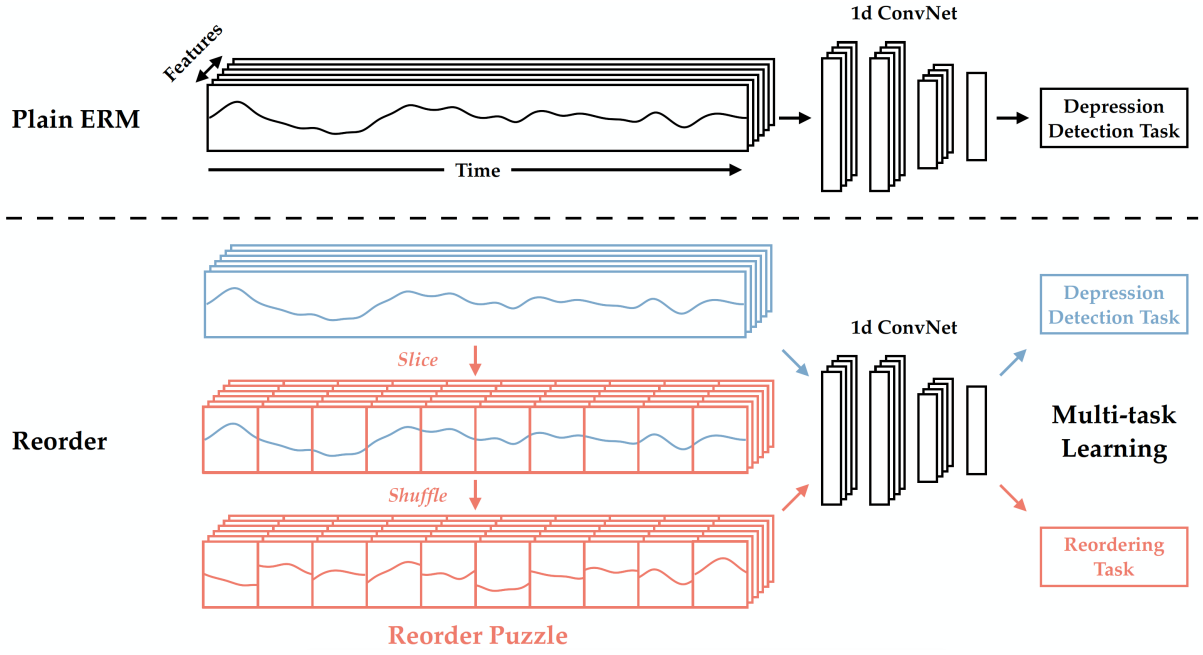


Fig. 2. The Design of Reorder Compared to ERM. In addition the main behavior modeling task, Reorder further introduces a secondary task of solving a reorder puzzle to force the model to learn the continuity of behavior trajectory.

trained via the following objective function:

$$\underset{\theta_f,\theta_c,\theta_r}{\operatorname{argmin}} \sum_{i=1}^{S} \Bigg(\underbrace{\sum_{j=1}^{N_i} \mathcal{L}_c(h(x_j^i|\theta_f,\theta_c), y_j^i)}_{\text{Loss Func of The Main Task}} + \underbrace{\sum_{j=1}^{\beta N_i} \alpha \mathcal{L}_r(h(z_j^i|\theta_f,\theta_r), p_j^i)}_{\text{Loss Func of The Reordering Task}} \Bigg)$$

where both $\mathcal{L}_c$ and $\mathcal{L}_r$ are cross-entropy losses. S is the total number of training domains, and N_i is the size of a domain i. α is used to control the weight of the reordering task while β is used to control the size of reordering data. $x_j^i, y_j^i, z_j^i, p_j^i$ are specific instances in each domain i with index j. Moreover, we also incorporate the Mixup augmentation technique [113] to increase the variation of the data. It is worth noting that the reorder puzzle is only enabled during the training stage. There is no shuffling at the testing stage to avoid extra noise.

4.4 Benchmark Platform

To gain a fair and reliable performance of these algorithms, we built a benchmark platform, **GLOBEM**, to incorporate all algorithms mentioned from Section 4.1 to 4.3. Compared to the existing platforms DomainBed [36] and DeepDG [94] that mainly aim for image-based domain generalization tasks, GLOBEM specifically focuses on longitudinal passive sensing data.

Figure 3 illustrates the overall structure of the GLOBEM platform. It splits the whole pipeline into three independent modules:

(1) The feature preparation module defines behavior features used by the algorithm;

(2) The model computation module defines how a behavior model is going to be trained. These two modules are determined by the core algorithm;

(3) The configuration module provides the flexibility to adjust hyperparameters in the algorithm.

Researchers and developers can re-use or re-purpose any of these modules to develop new algorithms within the pipeline. Moreover, GLOBEM separates the configuration setup from the model definition, supporting easy testing and ablation studies of hyperparameters and different features.

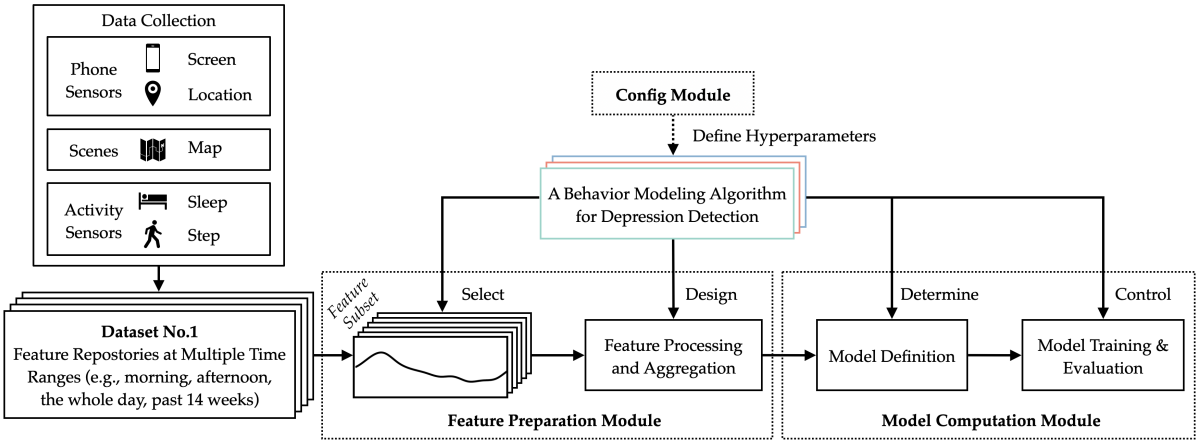


Fig. 3. Design of The Benchmark Platform GLOBEM. It modularizes the pipeline and supports flexible adjustment of the existing algorithms and easy development of new algorithms.

5 ANALYSIS

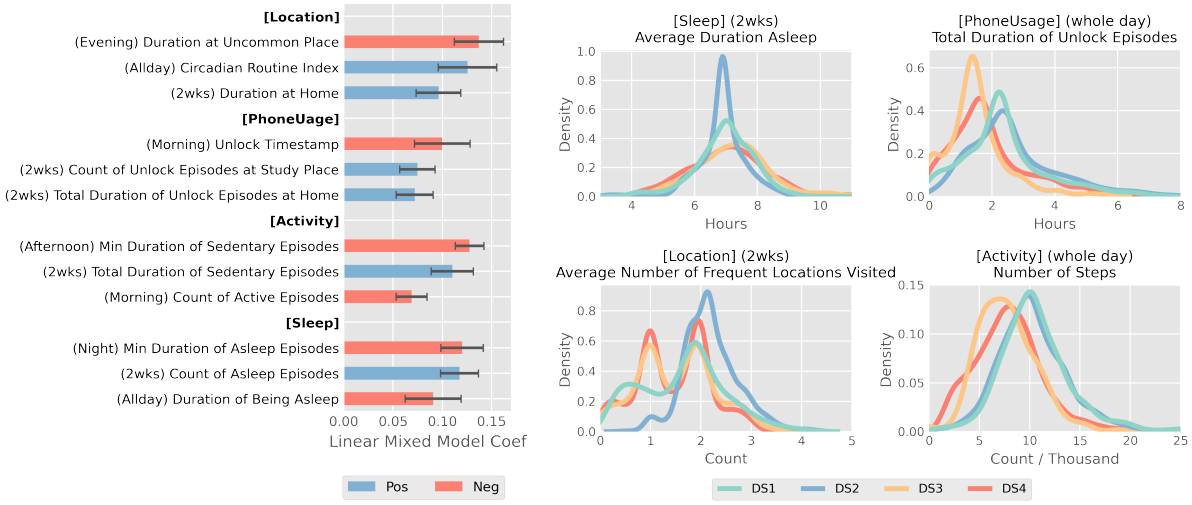
We summarize our analysis results in this section. We first reveal the feature difference among datasets (Section 5.1). We then demonstrate the significant performance drop between the existing depression detection models' performance on our datasets versus their reported results in previous literature (Section 5.2). We conducted cross-dataset evaluation to compare the generalizability of all methods, and demonstrated the advantage of our proposed *Reorder* (Section 5.3-5.6). Finally, we summarize our key findings in Section 5.7.

5.1 Dataset Analysis

We analyzed the commonalities and differences between the features in the four datasets (Section 5.1.1) Then, we conducted two tasks to quantify the distribution difference among the 1) datasets (Section 5.1.2) and 2) individuals (Section 5.1.3).

5.1.1 Feature Analysis. We started with the analysis between every single feature's value and the prediction target, *i.e.*, depression labels. We computed one linear mixed effect model for each feature (with participant ID as the random effect) within each dataset, and identified those significant features ($p < 0.05$) having the same directions in all datasets (*i.e.*, either all positive or all negative coefficient). Figure 4a summarizes representative features with large coefficients in each data type. For example, the sleep-related feature coefficients show that lower sleep duration and more interrupted sleep (as indicated by the count of asleep episodes) are associated with higher depression scores. We will discuss these results more in Section 6.

Figure 4b presents the distribution of some representative features of different data types. Some features have similar patterns. For example, the duration of being asleep has a peak around 7 hours, but the variance is different



(a) Linear Mixed Model Coefficients of Example Features

(b) Distribution Comparison of Example Features

Fig. 4. Features Analysis across All Datasets. (a) Each data type's top features with consistent coefficients of linear mixed effect models between the feature value and depression labels across all datasets. Red indicates negative coefficients and blue indicates positive coefficients. Error bar indicates standard error. (b) Example features' distribution across all datasets, which reveals how datasets can differ from each other. Datasets of the same institute are coded with closer colors. DS1 (green) and DS2 (blue) belong to the same institute, and DS3 (orange) and DS4 (red) belong to the other institute.

across datasets. Some features have similar patterns within the same institute. For example, DS1 and DS2 have closer distributions on the number of steps, while DS3 and DS4 have closer distributions, which could be related to location contexts of each institute. Meanwhile, some features have distinctive patterns. For example, the average number of frequent locations visited in DS1, DS3, and DS4 have two peaks, while only one peak is visible in DS2. These findings indicate that the same feature across datasets can often have various distributions.

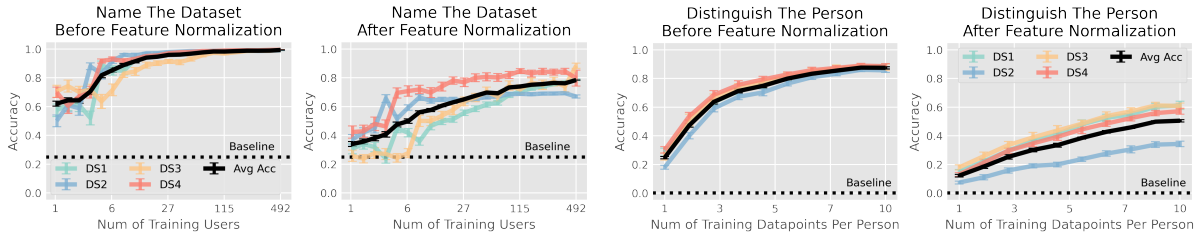
5.1.2 Domain Classification: Dataset as Domain. To quantify the difference among datasets, we first conducted a "Name-The-Dataset" task on our four datasets [86]. For every dataset, we first aggregated each feature matrix by calculating the mean along the time dimension to obtain a feature vector for each sample. We then performed an 80%/20% user split for the training/testing set, *i.e.*, no overlapping user's data is in both the training and testing sets simultaneously. We used a portion of the training data to train a small random forest model (10 decision trees, each with a maximum depth of 3) to classify which dataset a data belongs to (*i.e.*, four-class classification). We used SMOTE to mitigate the data imbalance as datasets have different sizes [17].

The left side of Figure 5a indicates the model performance on the testing set. With only 1 user in the training set (around 0.2% of samples from the training set), the model is able to achieve an average accuracy of 62.0%, compared to 25% as the baseline. With 5 users (1%) and 50 users (10%) from the training data, the accuracy reaches 81.7% and 96.8%, respectively. These results indicate that the behavior features from different datasets (*i.e.*, populations or years) have different distributions allowing them to be easily differentiated.

Feature normalization is one of the common techniques for mitigating feature value differences and aligning the data. Therefore, we also trained another model with normalized features (subtracting the median and divided by the 5-95 quantile range). As shown in the right side of Figure 5a, the model still achieves an accuracy of 34.0%, 49.9%. and 67.8% with 1, 5, and 50 users from the training set. The normalization does diminish the distribution shift, but the distinguishable differences between datasets still persists.

5.1.3 Domain Classification: Person as Domain. To quantify the differences among individuals, we further conducted a "Distinguish-The-Person" task, replacing the label from the dataset with the user ID. We performed an 80%/20% split on each user's data for the training/testing set. Similar to the "Name-The-Dataset" task, we then used a proportion of the training data to train another random forest model (maximum leaves=2k) to classify which user ID a data point belongs to, which is a challenging 534-class classification task (same as the total number of unique participants in the four datasets).

The two plots in Figure 5b show the performance using the features before and after normalization. Using 1, 5 and 10 data points per user from the training set, the model achieves 24.7%, 74.8%, 87.4% using direct features,



(a) "Name-The-Dataset" Task Results

(b) "Distinguish-The-Person" Task Results

Fig. 5. Results of Domain Classification Tasks using Simple Random Forest. Four colored lines indicate the accuracy of the four datasets, and the black line indicates the overall accuracy. Error bar indicates the standard error. The same below.

and 12.2%, 33.5%, 50.5% using normalized features, which are all significantly higher than the baseline accuracy 0.19% (1/534). We also tested the person-year classification (618 user-years), which showed similar results.

These analysis results clearly show that data from different datasets and individuals all have distinguishing distributions. In the "Distinguish-The-Person" task, the relative advantage over baseline is even larger than the "Name-The-Dataset" task, suggesting that the challenges of domain generalization in longitudinal human behavior modeling may come more from individual differences than dataset or population differences. We will present more results to support this claim in Section 5.4.

5.2 Prior Depression Detection Algorithms Performance Analysis

We followed the same procedure as the prior depression detection work to evaluate the performance of these 9 models on each of the four datasets. Some models focused on end-of-term detection [19, 27, 80, 100, 101], while others focused on frequent weekly detection [14, 55, 93, 95]. Thus we evaluated these models on both tasks. To keep the process consistent with the prior work, the models' hyperparameters were tuned via grid search with the same range as mentioned in each prior work. The training and testing used data from the same dataset, using twenty-fold cross-validation at the user level (*i.e.*, leaving 5% of the users out in each fold). For each of the four datasets, we repeated this process and calculated balanced accuracy as the metric, *i.e.*, the mean of sensitivity (true positive rate) and specificity (true negative rate). We picked this metrics for two reasons. First, most of the prior work reported metrics based on the confusion matrix (*e.g.*, precision, reall, F1 score), thus we can easily calculate and compare balanced accuracy from the prior work. Second, it has been shown more robust to class-imbalance compared to accuracy or F1 score [12].

Table 2 summarizes the results of model performance on each dataset. For the end-of-term depression detection, only three models, *Xu et al.- Interpretable*, *Xu et al.- Personalized*, and *Chikersal et al.*, achieve a balanced accuracy over 60%. However, these models' performance is significantly lower than the reported result in the prior work (average $\Delta = 15.9 \pm 10.7\%$). We have similar findings in the repeated weekly depression detection task. None of these models' performance reaches 60%. Again, we observe a big gap between the reported results and our results (average $\Delta = 22.6 \pm 8.5\%$). Figure 6 highlights the performance drop of these methods. Such findings indicate that prior algorithms do not generalize well on our datasets.

Table 2. Balanced Accuracy of Predicting End-of-term or Weekly Depression Status. Models are evaluated by 5-fold cross-validation within each of the four datasets. The "Prior Work" columns show the performance contrast between the reported results in prior literature (evaluated on their own datasets) and the results on our datasets. $^{+/++}$ means the literature only reported F1-score/ROC AUC, which was usually close to balanced accuracy.

Model	**Task1: End-of-Term Depression Detection**						**Task2: Weekly Depression Detection**					
	DS1	**DS2**	**DS3**	**DS4**	**Avg**	*Prior Work*	**DS1**	**DS2**	**DS3**	**DS4**	**Avg**	*Prior Work*
Majority Baseline	0.500	0.500	0.500	0.500	0.500	-	0.500	0.500	0.500	0.500	0.500	-
Canzian *et al.* [14]	0.559	0.516	0.526	0.500	0.525	-	0.512	0.497	0.512	0.500	0.505	*0.760*
Saeb *et al.* [80]	0.539	0.508	0.562	0.480	0.522	*0.791*	0.496	0.549	0.508	0.506	0.515	-
Farhan *et al.* [27]	0.552	0.609	0.505	0.620	0.572	*0.855*	0.519	0.515	0.519	0.513	0.517	-
Wahle *et al.* [93]	0.526	0.527	0.562	0.583	0.550	-	0.514	0.530	0.519	0.503	0.516	*0.616*
Lu *et al.* [55]	0.574	0.558	0.403	0.634	0.542	-	0.531	0.499	0.482	0.534	0.511	0.770^{+}
Wang *et al.* [98]	0.566	0.500	0.537	0.503	0.527	-	0.534	0.500	0.512	0.500	0.512	0.809^{++}
Xu *et al.*- Interpretable [100]	0.722	0.623	0.815	0.706	0.716	*0.806*	0.533	0.576	0.677	0.553	0.585	-
Xu *et al.*- Personalized [101]	0.723	0.699	0.818	0.649	0.722	*0.819*	0.568	0.562	0.614	0.572	0.579	-
Chikersal *et al.* [19]	0.728	0.776	0.795	0.698	0.749	*0.816*	0.615	0.613	0.595	0.551	0.593	-

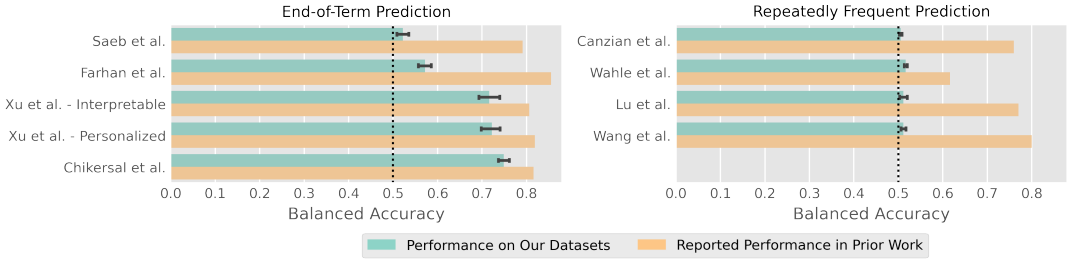


Fig. 6. The Performance Gap between Model Results on Our Datasets *v.s.* Reported Results in Prior Work, highlighting the results in Table 2. The dashed line indicates naive majority baseline.

5.2.1 Potential Explanation of Performance Drop. Since these algorithms were evaluated on separate datasets in previous literature, we speculate that there are two potential explanations.

First, although prior work has shown strong evidence that our included features of location, screen usage, physical activity, and sleep patterns already cover the most effective features from the literature [19, 100], the features excluded due to missing sensor data across datasets might also be important for a model to work effectively; Two prior algorithms (*Xu et al.- Interpretable* and *Xu et al.- Personalization*) were originally developed using DS1/DS2 raw data with a different feature extraction pipeline. Other than the slight difference on the user set (as mentioned in Section 4.1), the majority of the performance drop ($\Delta = 7.0\%$ and 6.7%, respectively) come from the missing features types, including Bluetooth, WiFi, call, message, and battery. Table 5 in the Appendix shows more detailed analysis of the performance difference. However, for some of the prior work (*e.g.*, *Saeb et al.* and *Canzian et al.*), we covered all the features in that original work and still observed a significant performance gap that is larger than *Xu et al.- Interpretable* and *Xu et al.- Personalization* ($\Delta > 20\%$), which leads to the next possible explanation:

Some features may be particularly tuned to specific datasets, which leads to feature overfitting to a certain group of participants in one dataset. These features cannot properly generalize for another group of participants from a different population and dataset. Thus, we conduct cross-dataset evaluation experiments to systematically measure the models' generalizability.

5.3 Cross-dataset Generalization Analysis

One of our core research goals is to explore the generalizability of behavior models. Using four datasets, we built a leave-one-dataset-out evaluation pipeline [73, 78]. Specifically, each time we took out one dataset as the testing set, and used the three other datasets as the training set. Such a cross-dataset analysis can effectively measure how a model trained on existing datasets could work on a new unseen dataset [50, 86].

5.3.1 Training Details. We focused on the weekly depression prediction as the main task, because the small sample size of the end-of-term task makes it infeasible to train deep models, *i.e.*, all of the models (17) other than the prior depression detection models (9). For the prior depression detection models, we followed a similar procedure as Section 5.2 to conduct a hyperparameter tuning by grid search on the three training datasets.

The rest of the methods (Section 4.2 and 4.3) all used deep models. For methods that used 1D-CNN as the backbone, we used a simple architecture based on a small-range tuning using ERM-1D-CNN: It had three 1D-convolution layers (size 8, stride 3, ReLU activation), each followed by a batch normalization layer, a max-pooling layer, as well as a dropout layer (rate 0.25). A fully connected layer (size 16) was attached after flattening the third

convolution layer's output to convert it into a vector of length 16. The following layers were then customized for each model.

For our new method *Reorder*, the feature layer was connected to two fully connected layers (size 16 and 2) for the classification task, and another two layers (size 32 and 200) for the reordering task. Additional model details are listed in codebase's config files. For the new method *Clustering*, we picked the cluster number as 60. The auto-encoder had two 1D-convolution layers (size 64 and 32) with the middle hidden size as 10. The Siamese network in each cluster adopted the same three 1D-convolution layers (size 8) architecture as other models.

Other architectures were also kept simple: 2D-CNN used three 2D-convolution layers with the same size, stride, and activation function as 1D-CNN; LSTM used two bi-directional layers (size 20); Transformer used two transformer blocks, each with 4 self-attention heads (size 4) and a 1D-convolutional feed forward layer (size 16).

For all models, we used Adam as the optimizer and adopted a cosine annealing schedule to repeatedly decrease the learning rate and then restart with a higher learning rate [53], with an initial learning rate of 0.001, an annealing decay of 0.95, and an annealing step size of 100. We isolated 10% of the training datasets as a validation set. All models are trained with 200 epochs, and the best epoch was picked based on the performance on the train and validation set. In addition to balanced accuracy, we further employed ROC AUC as the main evaluation metric as it indicates the overall results with varying decision boundaries in a detection problem.

5.3.2 Result Analysis. Table 3 lists out the results of all models on each of the four datasets, and Figure 7 presents the barplot of these models' ranked average performance. There are a few noteworthy observations.

First, all nine depression detection models have worse performance than that of Table 2. The best model, *Chikersal et al.*, has an average balanced accuracy of 52.0% in the cross-dataset evaluation, compared to 58.8% in the within-dataset evaluation. This performance gap, in addition to the previously described gap between the reported results in prior work and the results on our datasets, indicates that these models do not generalize well across datasets. Since all of our evaluation experiments use the same features, the first explanation of missing other sensors' features (at the end of Section 5.2) does not play a role at this stage. Therefore, the gap between within-dataset and cross-dataset evaluation is mainly caused by the distribution shift among different populations.

Second, modern ML techniques have been developed to deal with the challenge of feature shift across domains. However, these models barely work on our datasets. Among the 15 models we investigated, *CSD - Person as Domain* and *ERM - 2D-CNN* achieves the highest ROC AUC (52.3%). The results are similar to the results of traditional depression detection models (The best model *Chikersal et al.* achieves an ROC AUC of 54.1%.). These evaluation outcomes show that recent domain generalization methods do not work well on our datasets. This can be explained by the fact that most of these methods were developed under the context of CV or NLP tasks. Their generalizability may be affected when applied to longitudinal behavior data. Another interesting finding is the good performance of the naive ERM-based methods. Most of them (except *ERM-Transformer*) rank top 10 among the total of 26 methods, outperforming many generalization methods such as *DANN* and *MLDG*. Such a finding is consistent with the results in DomainBed [36] and DeepDG [94]. Both pointed out that ERM baseline often has competitive generalization performance.

Finally and most importantly, among all 26 models, our newly proposed *Reorder* model achieves the highest ROC AUC of 57.5% and the highest balanced accuracy of 55.2%. As shown in Figure 7, *Reorder* stands out among all methods. It outperforms the other models by at least 3.4% on ROC AUC (6.3% relative advantage), and 3.2% on absolute balanced accuracy (6.2% relative advantage), both with statistical significance ($p < 0.05$). Since *Reorder* has the same 1D-CNN backbone as *ERM-1D-CNN*, the comparison between these two models reveals the effect of adding the second reorder puzzle-solving task, which boosts the performance by 5.9% on ROC AUC (11.4% relative advantage) and 3.9% on balanced accuracy (7.6% relative advantage). Such an improvement illustrates that learning the temporal continuity of behavior trajectory can enhance the model's generalizability. However, although *Reorder* shows positive signals on domain generalization, it is worth noting that our model still has

Table 3. Model Performance of Predicting Weekly Depression Status across Datasets. Models are tested on one dataset after being trained on all other datasets. The Adv column indicates the advantage compared to the majority baseline. + or − indicates the algorithm has at least one or no metric better than the baseline, with t-test statistical significance: $p < 0.1_{\bullet}$ (marginal significance), $< 0.05^{*}$, $< 0.01^{**}$, and $< 0.001^{***}$.

Model	ROC AUC					Balanced Accuracy					Adv
	DS1	DS2	DS3	DS4	Avg	DS1	DS2	DS3	DS4	Avg	
Majority Baseline	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	
Canzian *et al.* [14]	0.490	0.493	0.457	0.537	0.494	0.499	0.500	0.500	0.500	0.500	−
Saeb *et al.* [80]	0.516	0.491	0.554	0.504	0.516	0.504	0.499	0.510	0.510	0.506	+
Farhan *et al.* [27]	0.509	0.472	0.483	0.493	0.489	0.517	0.489	0.509	0.489	0.501	+
Wahle *et al.* [93]	0.496	0.550	0.502	0.503	0.513	0.501	0.510	0.500	0.505	0.504	+
Lu *et al.* [55]	0.557	0.505	0.443	0.449	0.488	0.538	0.496	0.438	0.492	0.491	−
Wang *et al.* [98]	0.536	0.539	0.500	0.500	0.519	0.501	0.502	0.500	0.500	0.501	+
Xu *et al.*- Interpretable [100]	0.457	0.507	0.461	0.496	0.480	0.501	0.498	0.478	0.502	0.495	−
Xu *et al.*- Personalized [101]	0.482	0.534	0.548	0.517	0.520	0.484	0.530	0.511	0.516	0.510	+
Chikersal *et al.* [19]	0.590	0.525	0.526	0.523	0.541	0.545	0.503	0.523	0.508	0.520	$+^{*}$
ERM - 1D-CNN [89]	0.511	0.530	0.512	0.511	0.516	0.503	0.510	0.522	0.520	0.514	$+^{*}$
ERM - 2D-CNN [89]	0.508	0.509	0.535	0.542	0.523	0.507	0.504	0.523	0.534	0.517	$+^{*}$
ERM - LSTM [89]	0.518	0.516	0.536	0.511	0.520	0.511	0.502	0.528	0.502	0.511	$+^{*}$
ERM - Transformer [89]	0.508	0.464	0.527	0.486	0.496	0.513	0.474	0.510	0.488	0.496	−
ERM - Mixup [113]	0.512	0.522	0.513	0.520	0.517	0.517	0.505	0.519	0.513	0.514	$+^{***}$
IRM [2]	0.546	0.514	0.518	0.507	0.521	0.532	0.513	0.524	0.510	0.520	$+^{**}$
DANN - Dataset as Domain [32]	0.501	0.510	0.465	0.506	0.496	0.499	0.500	0.500	0.500	0.500	−
DANN - Person as Domain [32]	0.512	0.511	0.470	0.507	0.500	0.500	0.500	0.500	0.500	0.500	−
CSD - Dataset as Domain [76]	0.519	0.530	0.516	0.504	0.517	0.517	0.530	0.504	0.494	0.511	$+^{*}$
CSD - Person as Domain [76]	0.510	0.521	0.524	0.536	0.523	0.500	0.513	0.523	0.522	0.514	$+^{**}$
MLDG - Dataset as Domain [50]	0.495	0.475	0.519	0.496	0.496	0.501	0.470	0.523	0.501	0.499	−
MLDG - Person as Domain [50]	0.509	0.539	0.512	0.488	0.512	0.499	0.522	0.498	0.483	0.501	+
MASF - Dataset as Domain [26]	0.505	0.514	0.506	0.522	0.512	0.496	0.499	0.509	0.499	0.501	$+^{*}$
MASF - Person as Domain [26]	0.508	0.502	0.485	0.523	0.504	0.507	0.507	0.489	0.498	0.500	+
Siamese Network [47]	0.512	0.509	0.488	0.508	0.504	0.512	0.509	0.488	0.508	0.504	+
Clustering	0.522	0.502	0.497	0.521	0.511	0.518	0.505	0.499	0.517	0.510	$+_{\bullet}$
Reorder	**0.584**	**0.588**	**0.580**	**0.548**	**0.575**	**0.570**	**0.546**	**0.558**	**0.535**	**0.552**	$+^{***}$

great room for improvement. An ROC AUC of 57.5% is still far from being deployable in real-life scenarios, and we need more future research to improve model generalizability (see Section 6).

5.4 Cross-Institute and Cross-Year Generalization Analysis

In addition to the leave-one-dataset-out evaluation, we conducted additional experiments to obtain more insights into the models' generalizability and investigate different generalization challenges. As the four datasets were collected from two institutes across two years, we can evaluate how these models generalize across institutes (*i.e.*, different populations) (Section 5.4.1), and across years (*i.e.*, different users within the same population) (Section 5.4.2). Moreover, in each institute, there were a small number of people who participated in both years. Thus, we also evaluated the models on these subsets of users across years to test generalization across the same participants at different times (Section 5.4.3).

5.4.1 Cross-Institute. We used the two datasets from one institute as the training set, and the two datasets from the other institute as the testing set. Models' training setup was the same as Section 5.3.1. The left side of Figure 8

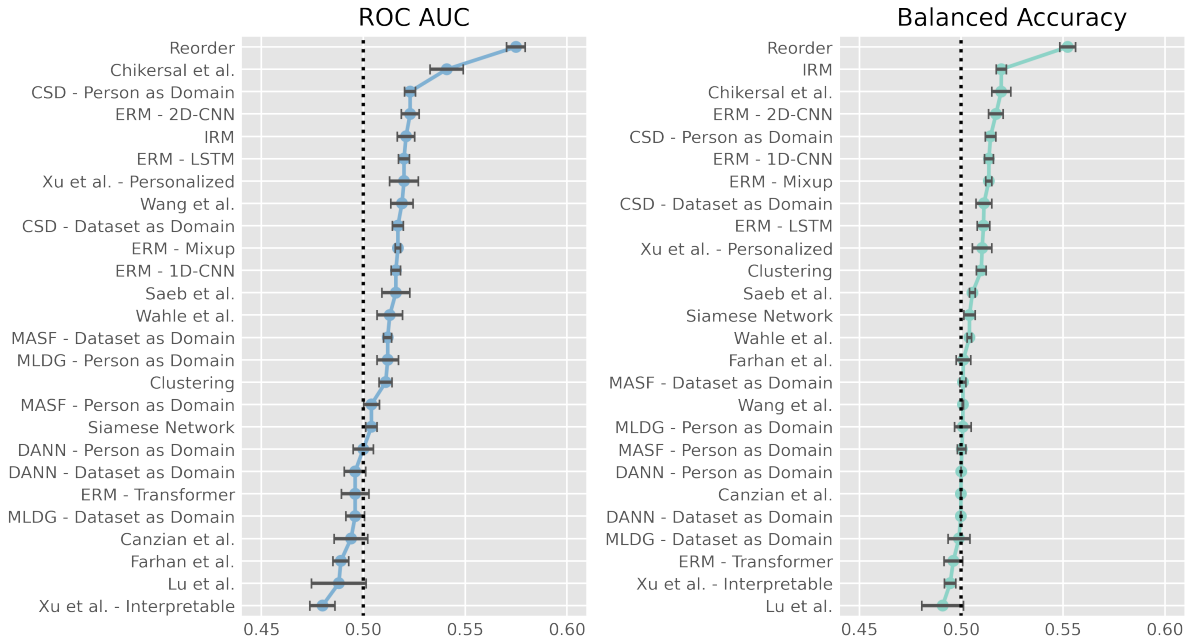


Fig. 7. Model Performance of Predicting Bi-Weekly Depression Status across Datasets. The models are ranked based on the average results in Table 3. The dashed line indicates naive majority baseline. The same below.

presents the ranked average balanced accuracy of all methods (train/test on institute 1/2 and then on institute 2/1). Overall, the performance is not as good compared to Figure 7. *Xu et al.- Personalized* has the best performance, but it only achieves an ROC AUC of 53.1%, followed by our method *Reorder* (52.4%). These lower performance results reflect that cross-institute generalization is a more challenging task.

5.4.2 Cross-Year. Similarly, we used two datasets from one year as the training set, and the rest from the other year as the testing set. The training setup was kept the same. The middle portion of Figure 8 shows the average balanced accuracy (train/test on year 1/2 and then on year 2/1). Compared to the cross-institute evaluation, the results of the cross-year evaluation are slightly better. Our method *Reorder* has the best performance, with an ROC AUC of 54.2%. This indicates that cross-year generalization could be easier than cross-institute generalization. Moreover, some models have interesting contrasting results. For example, *Clustering* ranks among the top 10 in the cross-institute evaluation, while it ranks among the bottom 5 in the cross-year evaluation (Δ = 2.1%). In contrast, *Xu et al.- Interpretable* has the worst performance in the leave-one-dataset-out evaluation, but ranks No.4 and No.9 in the cross-institute and cross-year evaluation (Δ = 4.0% and 3.9%). This means that these models capture different aspects of domain generalization.

5.4.3 Cross-Year with Overlapping People. We further narrowed down the evaluation to the overlapping people across years (there are no overlapping users across institutes), *i.e.*, we used trained a model on overlapping users in one dataset, and tested the model on these users in the other dataset, with the same training details and dataset setup as Section 5.4.2. The right side of Figure 8 shows the average balanced accuracy. We observe a strong increase in performance. *Reorder* again achieves the best performance with an ROC AUC of 61.6%, which is 7.4% higher than the best result in Section 5.4.2. The advantage could be explained by the fact that the same users'

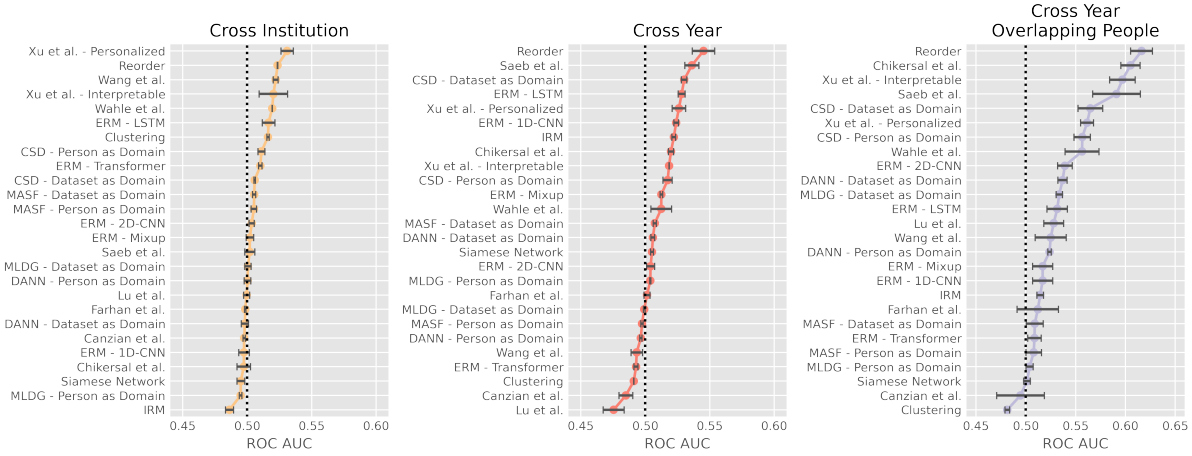


Fig. 8. Model Performance of Predicting Bi-Weekly Depression Status across Institutions (left) and Years (middle, right) . Models are tested on the datasets of one year/institution after being trained on the other year/institution.

behavior trajectories could preserve some patterns across years. In addition, both *Xu et al.- Interpretable* and *Xu et al.- Personalized* have good performance in the two cross-year evaluation, which is in line with their reported analysis [100, 101]. Moreover, *Chikersal et al.* ranks among top 5 among two of the four different setups. We speculate that the comprehensive feature aggregation and selection pipeline proposed in [19] may identify more generalizable features.

Overall, the cross-institute and cross-year evaluation results further illustrate more insights into model generalizability. Our model *Reorder* has the best or the second best results across the different tasks, revealing its advantages over other models. Moreover, the results of the third cross-dataset setup (*i.e.*, different times of the same users) are clearly better than those of the other two setups, which reveals that the individual differences (no matter whether that is within or between populations) may play the most important role in the cross-dataset generalization challenge.

5.5 Optimal Early Stopping Analysis

One of the major obstacles of domain generalization is the overfitting onto the training set [94, 118]. We also observed a similar overfitting issue during our deep models' training process. For example, during the training, *Reorder* achieves an average training ROC AUC of 74.9% and a training balanced accuracy of 67.7% for the main task, as well as an average training accuracy of 30.5% for the reordering task (as a 200-class classification task). These results of the main task are much better than the ones on the testing set (note that there is no reordering task on the testing set.). As the training epoch increases, we first observe a generally increasing performance curve on the testing set (learning), and then a decreasing trend (overfitting). If we can find the "optimal" training epoch for each model via early stopping, the results could then reflect the "upper bound" of these models' performance.

We conducted such an experiment with the similar leave-one-dataset-out setup as Section 5.3. Note that this experiment was only applicable to deep-learning-based algorithms as their training process has multiple epochs. Specifically, for every model, we iterated through the training epoch from 1 to 200. We performed the same epoch selection at each epoch number based on the best performance on the train and validation set. We then compared the test performance across the 200 epochs and identified the best epoch, assuming we could stop training at this epoch. Note that this step involved a little information leakage as it leveraged the testing set to select the epoch.

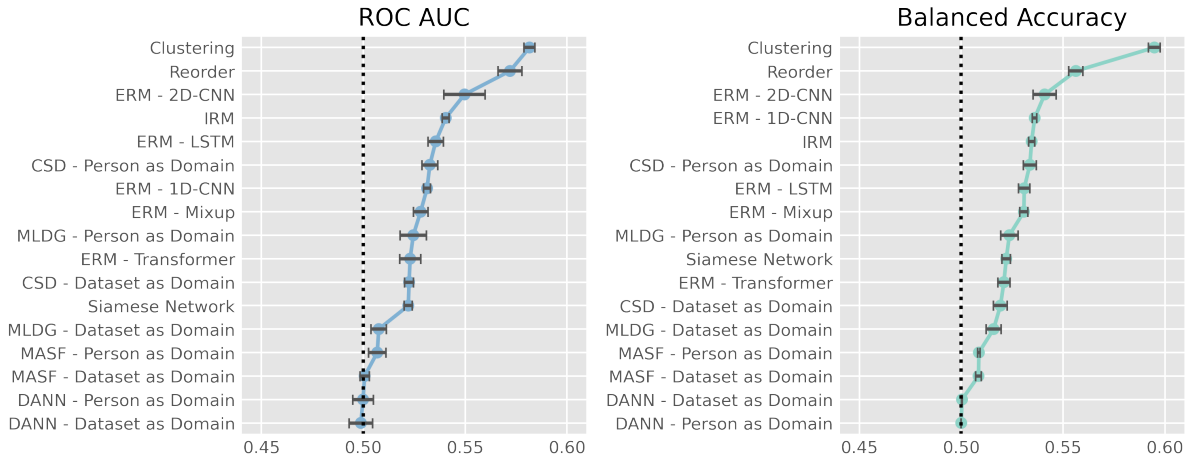


Fig. 9. Model Performance of Predicting Bi-Weekly Depression Status across Datasets. Models are tested on one dataset after being trained on all other datasets with the optimal epoch numbers.

Thus the results only reflect the theoretical upper bound performance. When we have a large validation dataset in the future, we could potentially achieve similar results as on the test set, so that information leakage will no longer be necessary.

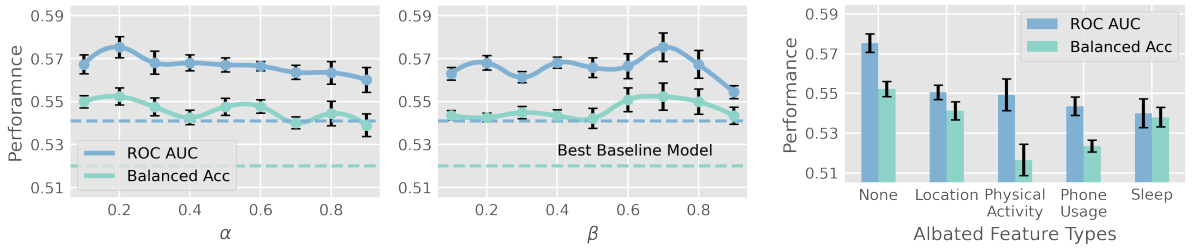
Figure 9 summarizes the generalizability of the deep-learning models to a new data set, when training is halted at the optimal training epoch. This figure shows that even under optimal conditions, current algorithms do not generalize well. In addition, we looked at *improvement over standard training*. *Clustering* achieves the best performance, with an average ROC AUC of 58.2% and an average balanced accuracy of 59.5%. This is a large benefit from optimal stopping (Δ equals 7.1% and 8.5%, respectively), which reveals that *Clustering* has an overfitting problem. *Reorder* achieves the second-best performance, but its performance improvement is minor ($\Delta < 0.5\%$). This indicates that there is minimal overfitting in *Reorder*.

5.6 Deeper Investigation of Reorder

We further conducted more analysis to obtain better understanding of *Reorder*, including additional hyper-parameter tuning experiments of the reordering task's loss function (Section 5.6.1) and an ablation study of different feature types (Section 5.6.2).

5.6.1 Reordering Task Tuning. The novelty of *Reorder* is mainly the secondary reordering task and the new loss function of the reorder task. To further investigate how much this task influences the results, we conducted experiments with different pairs of α (controlling the weight of the reordering task's loss) and β (controlling the amount of reordering data during the training process), using the leave-one-dataset-out setup throughout the experiments. Figure 10a presents the best results at different αs and βs.

Overall, the performance of *Reorder* is stable and consistently outperforms other baseline models. For α, the model performance increases as α increases before reaching 0.2, and then it slightly decreases with larger α. As for β, the model performance also generally increases before reaching 0.7, and then drops. Both results suggest that an appropriate weight and proportion of the reordering task is helpful, but an overemphasis on the secondary task (large weight, or large proportion) can hinder the performance of the main task.



(a) Tuning of Hyper-parameters

(b) Feature Ablation Study

Fig. 10. Extra Experiments on *Reorder* to Better Understanding the Model and Features. The experiments use the leave-one-dataset-out setup to ensure consistency. The dashed line indicates the performance of the best baseline models.

5.6.2 Feature Ablation Study. There are four feature types used in the model training: location, physical activity, phone usage, and sleep behaviors. In addition to knowing the overall performance of using all feature types, we are also interested in the importance of each type in the *Reorder* model. Therefore, we further conducted a feature ablation study by removing one of the four feature types from the feature set. All models are trained under the leave-one-dataset-out setup, and Figure 10b shows the results.

Unsurprisingly, removing any type of features would cause the model performance to drop. Based on ROC AUC, removing sleep-related features will cause the most significant performance drop ($\Delta = 3.5\%$), and removing location features will cause the least drop ($\Delta = 2.5\%$). Based on balanced accuracy, removing physical activity features will lead to the largest drop ($\Delta = 3.6\%$), and removing location features will lead to the least drop ($\Delta = 1.1\%$). These results suggest that compared to other feature types, location-related features could potentially be the least important for model generalizability in *Reorder*. This may be explained by the fact that different spatial contexts across datasets (from different institutions) can restrict on the generalizability of location features more than that of other feature types.

5.7 Results Highlight

To summarize, Table 4 highlights the key results of these models' performance under various setups.

For prior depression detection algorithms (using traditional machine learning), the "optimal" setup (the first column) has the results when an algorithm was trained and evaluated in a single dataset via 20-fold cross-validation. The performance differences between the optimal setup and cross-dataset setups (the right four columns) indicate that these models cannot generalize to other datasets very well.

For deep-learning-based domain generalization algorithms, the "optimal" setup has the results when a deep learning model could leverage the target dataset to select the best epoch, which is another perspective of "optimal" compared to the traditional machine learning methods [7]. The performance differences between the optimal setup and cross-dataset setups indicate that existing CV- or NLP-based domain generalization algorithms do not work well in the field of longitudinal behavior modeling. Although our new algorithms outperform these methods with statistical significance, they are still far from practical deployability and have great room for improvement.

Moreover, the comparison among the four cross-dataset setups indicates that cross-year generalization among the same users is an easier task (as reflected by the overall best performance in the last column), and that individual differences, either among the same institutes or across institutes, is more difficult and challenging for model generalization.

[7] The single dataset evaluation using deep-learning models has worse results compared to the best-epoch approach. We speculate the potential explanation is the limited sample size.

Table 4. ROC AUC Summary of Multiple Evaluation Setups on The Weekly Depression Detection Task. The first column indicates the "optimal" setup. For traditional depression detection models, it is 20-fold cross validation on single dataset. For deep learning models, it is the best epoch selected based on the target dataset.

Model	"Optimal" Setup Single Dataset	Leave-One-Dataset-Out	Cross Institute	Cross Year	Cross Year Overlapping
Canzian *et al.* [14]	0.525	0.494	0.498	0.487	0.495
Saeb *et al.* [80]	0.524	0.516	0.502	**0.530**	0.591
Farhan *et al.* [27]	0.572	0.489	0.499	0.501	0.512
Wahle *et al.* [93]	0.550	0.513	0.519	0.514	0.556
Lu *et al.* [55]	0.542	0.488	0.500	0.479	0.528
Wang *et al.* [98]	0.527	0.519	0.522	0.498	0.525
Xu *et al.*- Interpretable [100]	0.716	0.480	0.520	0.519	0.597
Xu *et al.*- Personalized [101]	0.722	0.520	**0.531**	0.527	0.562
Chikersal *et al.* [19]	**0.749**	**0.541**	0.497	0.516	**0.605**
Model	**"Optimal" Setup Best Epoch**	**Leave-One-Dataset-Out**	**Cross Institute**	**Cross Year**	**Cross Year Overlapping**
ERM - 1D-CNN [89]	0.531	0.516	0.498	0.520	0.517
ERM - 2D-CNN [89]	0.550	0.523	0.504	0.504	0.539
ERM - LSTM [89]	0.536	0.520	0.517	0.526	0.532
ERM - Transformer [89]	0.523	0.496	0.510	0.496	0.509
ERM - Mixup [113]	0.528	0.517	0.502	0.511	0.517
IRM [2]	0.541	0.521	0.486	0.516	0.515
DANN - Dataset as Domain [32]	0.499	0.496	0.498	0.505	0.537
DANN - Person as Domain [32]	0.500	0.500	0.500	0.497	0.524
CSD - Dataset as Domain [76]	0.522	0.517	0.506	0.526	0.565
CSD - Person as Domain [76]	0.533	0.523	0.511	0.516	0.557
MLDG - Dataset as Domain [50]	0.508	0.496	0.501	0.499	0.534
MLDG - Person as Domain [50]	0.525	0.512	0.495	0.502	0.504
MASF - Dataset as Domain [26]	0.501	0.512	0.505	0.507	0.509
MASF - Person as Domain [26]	0.507	0.504	0.505	0.499	0.508
Siamese Network [47]	0.522	0.504	0.495	0.504	0.501
Clustering	**0.582**	0.511	0.516	0.495	0.481
Reorder	0.572	**0.575**	**0.524**	**0.542**	**0.616**

6 DISCUSSION

In this section, we first discuss the relationship between our findings and previous depression literature (Section 6.1). We then analyze the challenges of domain generalization in our longitudinal behavior data format (Section 6.2), and discuss the potential alternative of domain adaptation besides domain generalization (Section 6.3). Finally, we summarize the limitations of our work (Section 6.4).

6.1 Evidence in Existing Depression-Related Literature

Our analysis in Section 5.1 identifies some features that are significantly associated with depression labels across all datasets. Compared to prior work that mainly used a single dataset, our multi-dataset analysis can reveal more generalizable findings at the feature level. For example, our sleep-related features show that participants with depression tended to have shorter sleep duration and more interrupted sleep. This is aligned with prior passive sensing studies [100, 101] and consistent with the criteria for diagnosing depression that sleep disturbance is a common symptom of depression [85, 87]. Our phone usage-related features indicate that participants checked

the phone more frequently (as indicated by the more count of and longer total duration screen unlock episodes). Such behavior was observed not only at home, but also at study places. Similar results were also observed in [80, 98]. This could reflect that participants with depression had more difficulties concentrating, another common symptom of depression [35, 56]. Moreover, other types of features further show that participants with depression tended to spend more time staying at home, remaining sedentary, and having fewer physical activities. These behaviors reflect a clear sign of a reduction of physical movement, which is one of the diagnosis criteria of depression [13, 79].

Interestingly, our location-related features also indicate that participants with a high depression scale score had a more consistent mobility routine. They visited less uncommon places than participants without depressive symptoms, and showed a stronger repetitive pattern in their locomotion trajectories. We did not find any depression literature that is directly related to this observation. We speculate that this lack of novelty seeking could be a sign of diminished interest in other activities other than the major routines (*e.g.*, going to class, having meals) – another common symptom of depression [3] – but it requires further analysis. Our findings may suggest some new research questions for behavioral science researchers.

6.2 Challenges of Domain Generalization

We observed a large gap between the performance of prior depression detection models across our 4 datasets and their reported performance in the corresponding literature on single datasets (Section 5.2). Combining the cross-dataset evaluation results of these models in Section 5.3, we found that the observed performance gap is mainly caused by the distribution shift between users in different datasets, *i.e.*, a feature that can be used to detect depression effectively in one dataset becomes less informative in another dataset. This observation highlights that depression detection work within a single dataset can potentially introduce dataset bias or feature distribution bias into model evaluation and conclusions. Moreover, the comparison of multiple cross-dataset setups in Section 5.4 further indicates that the individual behavior differences (both within and between populations) introduces a more significant domain generalization challenge than temporal differences of the same users (*i.e.*, how users may change from one year to the next).

However, such a dataset difference is not easy to address. Due to the limited amount of ground truth for each individual (a common property of longitudinal behavior datasets), generalization is more challenging. Our evaluation of recent deep learning domain generalization techniques indicates that these computer vision-based methods do not work well on longitudinal passive sensing data. Although we demonstrate that building a reordering puzzle can effectively force the model to learn behavior continuity, obtain generalizable behavior feature representation, and improve the generalizability under various setups, our *Reorder* model still cannot fully address the challenge of feature difference across datasets. It only achieved an ROC AUC of 57.5% and a balanced accuracy of 55.2% in the leave-one-dataset-out task. The model is not robust enough for real-life deployment, and we acknowledge that there still exists great room for improvement.

Our work serves as the first step toward cross-dataset model generalizability evaluation, and we hope future studies on behavior modeling algorithms will improve the generalizability across datasets and enable field deployment. Further, we provide an open-source platform on which other researchers and developers can conduct experiments and analyses about model generalizability.

6.3 Domain Adaptation and Model Personalization

Our analysis may also suggest the need to move from domain generalization to domain adaptation, *i.e.*, allow the model to access a small fraction of the data of new datasets or new users. As shown in our dataset analysis with the "Distinguish-The-Person" task (Figure 5b), a distribution shift exists even at the individual level. When trained on the same participants across years in both institutions, there is still a loss of performance, as reflected in the

right side of Figure 8. The best *Reorder* model's ROC AUC is below 65%. Thus, when treating each individual as a domain, a cross-dataset generalization would mean that a model needs to generalize to tens of (or even hundreds and thousands of) new domains/individuals in a new dataset, which could be overly challenging. A more feasible way is to allow the model to adapt and personalize to each user by tuning its parameters on a small amount of data from the target individual [10, 34, 52]. There have been some recent promising advances in this direction [33, 39]. Our analysis of optimal early stopping in Figure 9 also shows the potential of improving the model performance by accessing some data in the target dataset (using the target dataset as the test set to select the optimal epoch). If so, it provides the opportunity to leverage a wide range of transfer learning techniques in the domain adaptation area [74, 95]. This idea is practical in a real-life deployment: after a model is trained on existing datasets and applied to a new user, it can first accumulate data for the first few weeks and tune the model based on the new user's behavior. Such an adaptation process is necessary if we are moving towards the next step of providing personalized intervention experience [64, 108]. Although such a design would require the new user to provide extra behavior labels, it can potentially address the individual distribution shift and achieve model personalization.

6.4 Limitation and Future Work

There are a few limitations in our work. First, our current datasets only cover four data types and rely on RAPIDS for feature extraction. Due to the differences between the data collections of the two research groups, other data types (*e.g.*, Bluetooth, call logs, and audio) were not included. Also, our current datasets mainly have college students as participants, which may not represent the general population. We look forward to having more research groups join our generalization effort and contribute datasets that contain more data types and sub-populations. Moreover, the current feature set on RAPIDS is still limited. In the future, there are more potential behavioral features that can be integrated into RAPIDS. A second limitation is that our Institute1Year1 dataset does not contain a weekly well-established depression scale. Therefore, in Section 3.1.1, we trained a simple decision tree using Institute1Year2 dataset to generate labels for the Institute1Year1 dataset. Although the rule obtained from the decision tree (no depression if distressed is less than 2 and nervous is less than 3) is reasonable together with an acceptable accuracy of 74.5%, the rule might introduce errors that could distort the labels and the model training. Moreover, even the clinical measurement of depression is debated, making it inherently difficult to test models [30]. In the future, researchers may consider alternative ground truths as the prediction targets. Third, the current version of our platform GLOBEM implemented a limited number of domain generalization techniques. There are more techniques that seem promising, *e.g.*, CORAL [84], GroupDRO [81], ANDMask [75]. Moreover, our current algorithms mainly use cross entropy as the loss function, which is not designed for imbalanced datasets. Other loss functions that better deals with class imbalance, such as balanced cross entropy and focal loss [51], can be tested and evaluated on our platform. Our open-source platform was designed to be easily extensible and we plan include these algorithms in the future.

7 CONCLUSION

In this work, we highlight the importance of a behavior model's cross-dataset generalizability. Using depression detection as an example, we take the first step towards a systematic cross-dataset generalization evaluation in the longitudinal behavior modeling domain. We combined the efforts of two research groups across two institutes, each with two years of data, and established four datasets with a set of consistent features. We re-implemented nine prior depression detection methods, built eight recent domain generalization algorithms, and proposed two new methods for better generalizability. Our evaluation of these methods on our datasets demonstrated that existing algorithms barely outperform the baseline on cross-dataset generalization tasks, and that our new method *Reorder* could learn the continuity of behavior trajectories and achieve better generalizability across datasets.

Although statistically significant, its performance advantage is marginal, leaving great room for improvement. Moreover, the comparison of multiple generalization tasks indicates that individual behavior differences are the main source of challenges of domain generalization in the longitudinal behavior modeling area. We integrated all methods and open-sourced a benchmark platform named GLOBEM to assist future researchers in testing existing methods and developing new algorithms. We call for researchers in our field to pay more attention to the importance of cross-dataset generalizability evaluation. We envision this step as a necessary and essential part of behavior model deployment in the future.

ACKNOWLEDGMENTS

We thank Yasaman Sefidgar, Woo Suk Seo, Prerna Chikersal, Afsaneh Doryab for their contribution on the project. Yasaman Sefidgar and Woo Suk Seo significantly contributed to the data collection, data cleaning, initial feature extraction on a subset of Institute 1 datasets. Prerna Chikersal and Afsaneh Doryab kindly shared their source code to accelerate our implementation. Our studies were supported by the University of Washington (including the Paul G. Allen School of Computer Science and Engineering; Department of Electrical and Computer Engineering; Population Health; Addictions, Drug and Alcohol Institute; and the Center for Research and Education on Accessible Technology and Experiences); the National Science Foundation (EDA-2009977, CHS-2016365, CHS-1941537, IIS1816687 and IIS7974751), the National Institute on Disability, Independent Living and Rehabilitation Research (90DPGE0003-01), Samsung Research America, and Google.

REFERENCES

[1] Daniel A Adler, Fei Wang, David C Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE* (2022), 20.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]* (March 2020). http://arxiv.org/abs/1907.02893 arXiv: 1907.02893.

[3] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (dsm-5®)*. American Psychiatric Pub.

[4] Min S. Hane Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, John P. Pollak, Deborah Estrin, and Tanzeem Choudhury. 2016. Leveraging multi-modal sensing for mobile health: A case review in chronic pain. *IEEE Journal of Selected Topics in Signal Processing* 10, 5 (2016), 962–974.

[5] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung, and Anind K. Dey. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 5 (Jun 2017), 36 pages. https://doi.org/10.1145/3090051

[6] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–6.

[7] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K. Dey. 2016. Modeling and Understanding Human Routine Behavior. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2016), 248–260. https://doi.org/10.1145/2858036.2858557 ISBN: 9781450333627.

[8] Ian Barnett and Jukka-Pekka Onnela. 2020. Inferring mobility measures from GPS traces with missing data. *Biostatistics* 21, 2 (2020), e98–e112.

[9] Aaron T. Beck, Robert A. Steer, and Gregory K. Brown. 1996. Beck depression inventory-ii. *San Antonio* 78, 2 (1996), 490–498.

[10] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem Choudhury, Marta Hauser, et al. 2017. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric rehabilitation journal* 40, 3 (2017), 266.

[11] 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.

[12] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*. IEEE, 3121–3124.

[13] Terry C Camacho, Robert E Roberts, Nancy B Lazarus, George A Kaplan, and Richard D Cohen. 1991. Physical activity and depression: evidence from the Alameda County Study. *American journal of epidemiology* 134, 2 (1991), 220–231.

[14] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015), 1293–1304.

https://doi.org/10.1145/2750858.2805845
[15] Liangliang Cao, Zicheng Liu, and Thomas S Huang. 2010. Cross-dataset action detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1998–2005.
[16] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain Generalization by Solving Jigsaw Puzzles. *arXiv:1903.06864 [cs]* (April 2019). http://arxiv.org/abs/1903.06864 arXiv: 1903.06864.
[17] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[18] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 145–152.
[19] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G. Creswell, Jennifer Mankoff, J. David Creswell, Mayank Goel, and Anind K. Dey. 2021. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing. *ACM Transactions on Computer-Human Interaction* 28, 1 (Jan. 2021), 1–41. https://doi.org/10.1145/3422821
[20] Yucel Cimtay and Erhan Ekmekcioglu. 2020. Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition. *Sensors* 20, 7 (2020), 2034.
[21] Taylor Cox. 1994. *Cultural diversity in organizations: Theory, research and practice*. Berrett-Koehler Publishers.
[22] John R Crawford and Julie D Henry. 2004. The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology* 43, 3 (2004), 245–265.
[23] Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, 256–263. https://www.aclweb.org/anthology/P07-1033
[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). http://arxiv.org/abs/1810.04805 arXiv: 1810.04805.
[25] Afsaneh Doryab, Prerna Chikarsel, Xinwen Liu, and Anind K Dey. 2018. Extraction of behavioral features from smartphone and wearable data. *arXiv preprint arXiv:1812.10394* (2018).
[26] Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain Generalization via Model-Agnostic Learning of Semantic Features. *arXiv:1910.13580 [cs]* (Oct. 2019). http://arxiv.org/abs/1910.13580 arXiv: 1910.13580.
[27] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE Wireless Health (WH)*. IEEE, 1–8. https://doi.org/10.1109/WH.2016.7764553
[28] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28, 4 (2006), 594–611.
[29] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: Mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6.
[30] Eiko I Fried, Jessica K Flake, and Donald J Robinaugh. 2022. Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology* (2022), 1–11.
[31] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2017. Domain-Adversarial Training of Neural Networks. In *Domain Adaptation in Computer Vision Applications*. Springer International Publishing, Cham, 189–209. https://doi.org/10.1007/978-3-319-58347-1_10 Series Title: Advances in Computer Vision and Pattern Recognition.
[32] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2017. Domain-Adversarial Training of Neural Networks. In *Domain Adaptation in Computer Vision Applications*, Gabriela Csurka (Ed.). Springer International Publishing, 189–209. https://doi.org/10.1007/978-3-319-58347-1_10 Series Title: Advances in Computer Vision and Pattern Recognition.
[33] Taesik Gong, Yewon Kim, Adiba Orzikulova, Yunxin Liu, Sung Ju Hwang, Jinwoo Shin, and Sung-Ju Lee. 2021. DAPPER: Performance Estimation of Domain Adaptation in Mobile Sensing. *arXiv preprint arXiv:2111.11053* (2021).
[34] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. MetaSense: few-shot adaptation to untrained conditions in deep mobile sensing. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 110–123.
[35] Ian H Gotlib and Jutta Joormann. 2010. Cognition and depression: current status and future directions. *Annual review of clinical psychology* 6 (2010), 285–312.
[36] Ishaan Gulrajani and David Lopez-Paz. 2021. In Search of Lost Domain Generalization. (2021), 29.
[37] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[38] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. 2017. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*. Springer, 373–382.

[39] Joy He-Yueya, Benjamin Buck, Andrew Campbell, Tanzeem Choudhury, John M Kane, Dror Ben-Zeev, and Tim Althoff. 2020. Assessing the relationship between routine and schizophrenia symptoms with passively sensed measures of behavioral stability. *NPJ schizophrenia* 6, 1 (2020), 1–8.

[40] Geoffrey M Hodgson. 1997. The ubiquity of habits and rules. *Cambridge journal of economics* 21, 6 (1997), 663–684.

[41] Steven Hoffman, Renu Sharma, and Arun Ross. 2018. Convolutional neural networks for iris presentation attack detection: Toward cross-dataset and cross-sensor generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1620–1628.

[42] Xiao Hu and Yi-Hsuan Yang. 2017. Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs. *IEEE Transactions on Affective Computing* 8, 2 (2017), 228–240.

[43] Jeremy F Huckins, Alex W DaSilva, Elin L Hedlund, Eilis I Murphy, Courtney Rogers, Weichen Wang, Mikio Obuchi, Paul E Holtzheimer, Dylan D Wagner, and Andrew T Campbell. 2020. Causal factors of anxiety and depression in college students: longitudinal ecological momentary assessment and causal analysis using Peter and Clark momentary conditional independence. *JMIR mental health* 7, 6 (2020), e16684.

[44] Jeremy F Huckins, Alex W DaSilva, Weichen Wang, Elin Hedlund, Courtney Rogers, Subigya K Nepal, Jialing Wu, Mikio Obuchi, Eilis I Murphy, Meghan L Meyer, et al. 2020. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: Longitudinal smartphone and ecological momentary assessment study. *Journal of medical Internet research* 22, 6 (2020), e20185.

[45] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5001–5009.

[46] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. 2021. SelfReg: Self-Supervised Contrastive Regularization for Domain Generalization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2021), 10.

[47] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. *Proceedings of the 32nd International Conference on Machine Learning* (2015), 8.

[48] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.

[49] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (2010).

[50] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2017. Learning to Generalize: Meta-Learning for Domain Generalization. *arXiv:1710.03463 [cs]* (Oct. 2017). http://arxiv.org/abs/1710.03463 arXiv: 1710.03463.

[51] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[52] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. 2021. MetaPhys: few-shot adaptation for non-contact physiological measurement. In *Proceedings of the Conference on Health, Inference, and Learning*. ACM, Virtual Event USA, 154–163. https://doi.org/10.1145/3450439.3451870

[53] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[54] Bernd Löwe, Inka Wahl, Matthias Rose, Carsten Spitzer, Heide Glaesmer, Katja Wingenfeld, Antonius Schneider, and Elmar Brähler. 2010. A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of affective disorders* 122, 1-2 (2010), 86–95.

[55] Jin Lu, Jinbo Bi, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, and Bing Wang. 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. https://doi.org/10.1145/3191753 ISBN: 9781450351980.

[56] Donatella Marazziti, Giorgio Consoli, Michela Picchetti, Marina Carlini, and Luca Faravelli. 2010. Cognitive impairment in major depression. *European journal of pharmacology* 626, 1 (2010), 83–86.

[57] Stephen M Mattingly, Julie M Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K D'Mello, Anind K Dey, et al. 2019. The Tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[58] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. 2021. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine* 13, 586 (2021), eabb1655.

[59] Frances J Milliken and Luis L Martins. 1996. Searching for common threads: Understanding the multiple effects of diversity in organizational groups. *Academy of management review* 21, 2 (1996), 402–433.

[60] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. 2014. Toss "n" turn: Smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 477–486. https://doi.org/10.1145/2556288.2557220

[61] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino G. Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, et al. 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2 (2019), 37:1–37:24. https://doi.org/10.1145/3328908

[62] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching Hua Chen, and David Kotz. 2020. Evaluating the reproducibility of physiological stress detection models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020).

[63] Tom M Mitchell and Tom M Mitchell. 1997. *Machine learning*. Vol. 1. McGraw-hill New York.

[64] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine* 52, 6 (May 2018), 446–462. https://doi.org/10.1007/s12160-016-9830-8

[65] Ebrahim Nemati, Xuhai Xu, Viswam Nathan, Korosh Vatanparvar, Tousif Ahmed, Md Mahbubur Rahman, Dan McCaffrey, Jilong Kuang, and Alex Gao. 2022. Ubilung: Multi-Modal Passive-Based Lung Health Assessment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 551–555.

[66] Subigya Nepal, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Alex daSilva, Meghan Meyer, and Andrew Campbell. 2022. COVID Student Study: A Year in the Life of College Students during the COVID-19 Pandemic Through the Lens of Mobile Phone Sensing. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 42, 19 pages. https://doi.org/10.1145/3491102.3502043

[67] Daniel Nettle. 2006. The evolution of personality variation in humans and other animals. *American Psychologist* 61, 6 (2006), 622.

[68] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. 2017. Exploring generalization in deep learning. *Advances in neural information processing systems* 30 (2017).

[69] Stefanie Nickels, Matthew D Edwards, Sarah F Poole, Dale Winter, Jessica Gronsbell, Bella Rozenkrants, David P Miller, Mathias Fleck, Alan McLean, Bret Peterson, et al. 2021. Toward a mobile platform for real-world digital measurement of depression: User-centered design, data quality, and behavioral and clinical modeling. *JMIR mental health* 8, 8 (2021), e27589.

[70] Shirin Nilizadeh, Hojjat Aghakhani, Eric Gustafson, Christopher Kruegel, and Giovanni Vigna. 2019. Think Outside the Dataset: Finding Fraudulent Reviews using Cross-Dataset Analysis. (2019), 3108–3115. https://doi.org/10.1145/3308558.3313647 ISBN: 9781450366748.

[71] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*. Springer, 69–84.

[72] NSDUH 2018. the national survey on drug use and health - survey report in 2018. https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHNationalFindingsReport2018/NSDUHNationalFindingsReport2018.pdf.

[73] Jaya L. Padmanabhan, Danielle Cooke, Juho Joutsa, Shan H. Siddiqi, et al. 2019. A Human Depression Circuit Derived From Focal Brain Lesions. *Biological Psychiatry* 86, 10 (2019), 749–758. https://doi.org/10.1016/j.biopsych.2019.07.023 Cortical Pathology and Depression.

[74] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 22, 10 (2010), 1345–1359.

[75] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. 2020. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329* (2020).

[76] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. 2020. Efficient Domain Generalization via Common-Specific Low-Rank Decomposition. *arXiv:2003.12815 [cs, stat]* (April 2020). http://arxiv.org/abs/2003.12815 arXiv: 2003.12815.

[77] Alan Ramponi and Barbara Plank. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. *arXiv:2006.00632* (2020).

[78] Markus Riester, Wei Wei, Levi Waldron, Aedin C Culhane, , et al. 2014. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *Journal of the National Cancer Institute* 106, 5 (2014), dju048.

[79] Babak Roshanaei-Moghaddam, Wayne J Katon, and Joan Russo. 2009. The longitudinal effects of depression on physical activity. *General hospital psychiatry* 31, 4 (2009), 306–315.

[80] Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research* 17, 7 (2015), 1–11. https://doi.org/10.2196/jmir.4273

[81] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).

[82] Yasaman S. Sefidgar, Woosuk Seo, Kevin S. Kuehn, Tim Althoff, Anne Browning, Eve Riskin, Paula S. Nurius, Anind K. Dey, and Jennifer Mankoff. 2019. Passively-sensed behavioral correlates of discrimination events in college students. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 114 (Nov 2019), 29 pages. https://doi.org/10.1145/3359216

[83] Keith E Stanovich and Richard F West. 1998. Individual differences in rational thought. *Journal of experimental psychology: general* 127, 2 (1998), 161.

[84] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, 443–450.

[85] Michael E. Thase. 1998. Depression, sleep, and antidepressants. *The Journal of Clinical Psychiatry* (1998).

[86] Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, Vol. 2011. IEEE, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347 Issue: 28.

[87] Norifumi Tsuno, Alain Besset, and Karen Ritchie. 2005. Sleep and depression. *The Journal of Clinical Psychiatry* (2005).

[88] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*. 4068–4076.

[89] Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* 10, 5 (1999), 988–999.

[90] Julio Vega, Beth T Bell, Caitlin Taylor, Jue Xie, Heidi Ng, Mahsa Honary, Roisin McNaney, et al. 2022. Detecting Mental Health Behaviors Using Mobile Interactions: Exploratory Study Focusing on Binge Eating. *JMIR mental health* 9, 4 (2022), e32146.

[91] Julio Vega, Meng Li, Kwesi Aguillera, Nikunj Goel, Echhit Joshi, Kirtiraj Khandekar, Krina C Durica, Abhineeth R Kunta, and Carissa A Low. 2021. Reproducible Analysis Pipeline for Data Streams: Open-Source Software to Process Data Collected With Mobile Devices. *Frontiers in Digital Health* 3 (2021).

[92] Theo Vos and the GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. 2016. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the global burden of disease study 2015. *The Lancet* 388, 10053 (2016), 1545–1602.

[93] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. *JMIR mHealth and uHealth* 4, 3 (2016), e111. https://doi.org/10.2196/mhealth.5960 ISBN: doi:10.2196/mhealth.5960.

[94] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. 2021. Generalizing to Unseen Domains: A Survey on Domain Generalization. *arXiv:2103.03097 [cs]* (Dec. 2021). http://arxiv.org/abs/2103.03097 arXiv: 2103.03097.

[95] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153. https://doi.org/10.1016/j.neucom.2018.05.083 Publisher: Elsevier B.V..

[96] 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*. ACM, 3–14.

[97] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 295–306.

[98] 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. https://doi.org/10.1145/3191775 ISBN: 2474-9567.

[99] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised Deep Embedding for Clustering Analysis. *Proceedings of the 33 rd International Conference on Machine Learning* (2016), 10.

[100] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 1–33. https://doi.org/10.1145/3351274

[101] Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–27. https://doi.org/10.1145/3448107

[102] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3501904

[103] Xuhai Xu, Ahmed Hassan Awadallah, Susan T. Dumais, Farheen Omar, Bogdan Popp, Robert Rounthwaite, and Farnaz Jahanbakhsh. 2020. Understanding User Behavior For Document Recommendation. In *Proceedings of The Web Conference 2020*. ACM, Taipei Taiwan, 3012–3018. https://doi.org/10.1145/3366423.3380071

[104] Xuhai Xu, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021. Listen2Cough: Leveraging End-to-End Deep Learning Cough Detection Model to Enhance Lung Health Assessment Using Passively Sensed Audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–22. https://doi.org/10.1145/3448124

[105] Xuhai Xu, Haitian Shi, Xin Yi, Wenjia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing*

Systems. ACM, Honolulu HI USA, 14.
[106] Xuhai Xu, Chun Yu, Yuntao Wang, and Yuanchun Shi. 2020. Recognizing Unintentional Touch on Interactive Tabletop. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (March 2020), 1–24. https://doi.org/10.1145/3381011
[107] Xuhai Xu, Han Zhang, Yasaman S Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Scott Kuehn, Mike A Merrill, Paula S Nurius, Shwetak Patel, Tim Althoff, Margaret E Morris, Eve A. Riskin, Jennifer Mankoff, and Anind Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=GKOa7yNH8Uh
[108] Xuhai Xu, Tianyuan Zou, Han Xiao, Yanzhang Li, Ruolin Wang, Tianyi Yuan, Yuntao Wang, Yuanchun Shi, Jennifer Mankoff, and Anind K Dey. 2022. TypeOut: Leveraging Just-in-Time Self-Affirmation for Smartphone Overuse Reduction. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–17. https://doi.org/10.1145/3491102.3517476
[109] Runze Yan, Whitney R Ringwald, Julio Vega, Madeline Kehl, Sang Won Bae, Anind K Dey, Carissa A Low, Aidan GC Wright, and Afsaneh Doryab. 2022. Exploratory machine learning modeling of adaptive and maladaptive personality traits from passively sensed behavior. *Future Generation Computer Systems* 132 (2022), 266–281.
[110] Jenny Yang, Andrew AS Soltan, and David A Clifton. 2022. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *npj Digital Medicine* 5, 1 (2022), 1–8.
[111] Yi Yao and Gianfranco Doretto. 2010. Boosting for transfer learning with multiple sources. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1855–1862.
[112] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation* 26, 2 (2007), 289–315.
[113] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412 [cs, stat]* (April 2018). http://arxiv.org/abs/1710.09412 arXiv: 1710.09412.
[114] Han Zhang, Margaret E. Morris, Paula S. Nurius, Kelly Mack, Jennifer Brown, Kevin S. Kuehn, Yasaman S. Sefidgar, Xuhai Xu, Eve A. Riskin, Anind K. Dey, and Jennifer Mankoff. 2022. Impact of Online Learning in the Context of COVID-19 on Undergraduates with Disabilities and Mental Health Concerns. *ACM Transactions on Accessible Computing* (July 2022), 3538514.
[115] Jing Zhang, Wanqing Li, Philip Ogunbona, and Dong Xu. 2019. Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition. *Comput. Surveys* 52, 1 (Feb. 2019), 1–38. https://doi.org/10.1145/3291124
[116] Xu Zhang, Junghyun Kim, Qingwei Lin, Keunhak Lim, Shobhit O Kanaujia, Yong Xu, Kyle Jamieson, Aws Albarghouthi, Si Qin, Michael J Freedman, et al. 2019. Cross-dataset time series anomaly detection for cloud systems. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*. 1063–1076.
[117] Yi Zhang and Rui Wang. 2009. Cross-domain dependency parsing using a deep linguistic grammar. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 378–386.
[118] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2021. Domain Generalization in Vision: A Survey. *arXiv:2103.02503 [cs]* (July 2021). http://arxiv.org/abs/2103.02503 arXiv: 2103.02503.

APPENDIX: ABLATION STUDY RESULTS

Table 5. Ablation Analysis on DS1 of Two Algorithms Developed on Overlapping Datasets. User set difference: consistent inclusion criteria cross all other algorithm *v.s.* more strict data filtering criteria with fewer users; Evaluation difference: 20-fold cross validation *v.s.* leave-one-user-out; Feature Difference: four feature types *v.s.* additional location contexts, Bluetooth, WiFi, call, battery features.

Changes	**Balanced Accuracy**	
	Xu *et al.*- Interpretable [100]	Xu *et al.*- Personalized [101]
Our Results	0.722	0.738
Remove User Set Difference	0.719	0.748
Remove Evaluation Difference	0.749	0.742
Remove Feature Difference	0.792	0.805
Original Reported Results	0.806	0.819