Xuhai (Orson) Xu *MIT Cambridge, MA* **Xin Liu** *Google Consumer Health Research, and University of Washington, Seattle, WA*
Han Zhang *University of Washington, Seattle, WA* **Weichen Wang** *Meta*
Subigya Nepal *Dartmouth College, Hanover, NH* **Yasaman S. Sefidgar** *University of Washington, Seattle, WA*
Woosuk Seo *University of Michigan School of Information, Ann Arbor, MI*
Kevin S. Kuehn *Department of Medicine at the University of California, San Diego, CA*
Jeremy F. Huckins *Bicogniv Inc., Burlington, VT* **Margaret E. Morris** *University of Washington, Seattle, WA*
Paula S. Nurius *University of Washington School of Social Work, Seattle, WA*
Eve A. Riskin *Stevens Institute of Technology, Hoboken, NJ* **Shwetak Patel** *University of Washington, Seattle, WA*
Tim Althoff, Andrew T. Campbell *Dartmouth College, Hanover, NH* **Anind K. Dey** *University of Washington, Seattle, WA*
Jennifer Mankoff *University of Washington, Seattle, WA*

Illustration, istockphoto.com

GLOBEM: CROSS-DATASET GENERALIZATION OF LONGITUDINAL HUMAN BEHAVIOR MODELING

Excerpted from "GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling," from IMWUT 2023: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* with permission. https://dl.acm.org/doi/10.1145/3569485 ©ACM 2023

Ubiquitous computing, the seamless integration of sensing, analytics, and feedback into daily life envisioned by Weiser [12], has come closer to reality with the broad adoption of smartphones and wearable devices. These devices, integral to users' daily routines, passively collect massive amounts of data on human behavior, offering unprecedented insights into personal health and well-being [7]. For example, passive sensing can continuously monitor subtle changes in behavior indicative of depression or other shifts in mental health status [10,14,15].

However, the challenge lies in the generalization of behavior models across diverse datasets, which often reflect different populations or conditions. Most models are typically trained and validated on data from a single source, limiting their applicability to broader populations or real-world deployment.

Our study addresses this challenge by examining the generalizability of longitudinal behavior models across multiple datasets, using depression detection as an example application. In this work, we evaluate the robustness of models from prior work and introduce a novel algorithm, *Reorder*, that leverages the behavioral science insights of temporal continuity and enhances model generalizability [13].

We further contribute to the field by presenting *GLOBEM* [13], an open-source benchmark platform that consolidates a number of algorithms to foster open-source research and development in this area. This platform allows for rigorous evaluation across multiple datasets and health prediction targets. It also supports flexible extension of new algorithms, new datasets, and new prediction targets.

This work underscores the importance of cross-dataset validation. We provide a comprehensive framework for other researchers to evaluate and enhance their behavioral models. Through this collective effort, we aim to ensure models are robust and applicable across different demographic and temporal contexts and across various health concerns. Figure 1 highlights our contributions.

MULTI-YEAR MULTI-INSTITUTION DATA COLLECTION

Our research builds upon large-scale longitudinal passive sensing data collected from smartphones and wearable devices. This data captures daily behavioral signals that are used for prediction model development, such as physical activity levels, social interactions, and mental health states.

Data Collection

Our collaborative study involved two research groups from two Carnegie-classified R-1 universities in the States. At each university, we conducted two longitudinal passive sensing data collection studies in two consecutive years, generating four datasets [7]. The collection studies followed a similar design to ensure sensor consistency across cohorts, and each group followed a uniform data transformation and feature extraction process, creating four institute-year datasets.

Participants in these studies were undergraduate students and were compensated based on their compliance with the study protocol. Our studies were approved by university institutional review boards (IRBs). The data collection included various modalities such as location, phone usage, physical activity, and sleep. In both institutions, we employed well-established and validated questionnaires to assess depression and other mental distress on a weekly basis and at the end of the quarter. The weekly surveys included PANAS (Institute1Year1 only), PHQ-4 (remaining datasets). The end-term surveys included BDI-II (Institute1) and PHQ-4 (Institute2).

As an initial step of model generalizability evaluation, we focused on a binary classification task: the presence of at least mild depressive symptoms as self-reported (i.e., PHQ-4 > 2, BDI-II > 13). Note that the PANAS contains questions related to depressive symptoms (e.g., "distressed") but does not assess depression in the style of measures such as the PHQ-4 or BDI-II. We had PHQ-4 and BDI-II scores for all datasets except Institute1Year1. To make the datasets compatible, we generate reliable ground truth binary labels from PANAS by developing a simple decision tree model using the Institute1Year2 dataset, which has both PANAS and PHQ-4 scores. The model achieves an accuracy of 74.5%. We then applied this model to Institute1Year1 dataset to generate labels. A comprehensive breakdown of study information and participant demographics can be found in Figure 2.

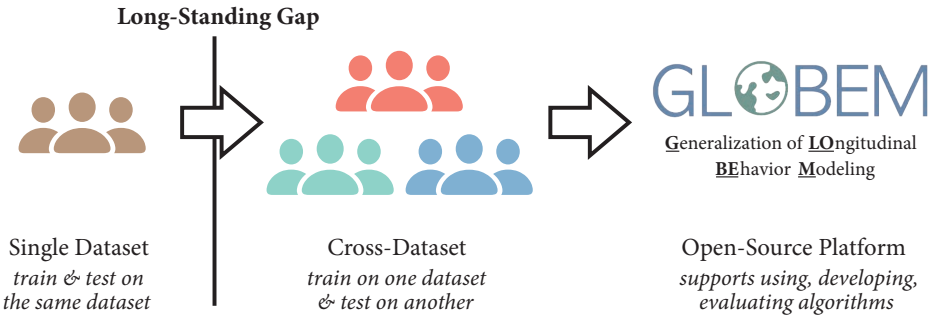


FIGURE 1. Overview of The Contributions of This Work. We systematically evaluate cross-dataset generalizability of 18 algorithms: 9 prior behavior modeling algorithms for depression detection, 8 recent domain generalization algorithms, and the new algorithm, Reorder, proposed in this paper. Our open-source platform GLOBEM consolidates these algorithms and provides support to researchers using, developing, or evaluating various algorithms.

	Institute 1		Institute 2	
	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, N4, W 100, B&H 1, NA 13 - 58 also in Year1
Gound Truth	- Weekly: Depression & Affect (44.4%) - End-term: BDI-II (35.4%)	- Weekly: PHQ-4 (50.3%) - End-term: BD-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	

FIGURE 2. Basic Study Information and Participant Demographics of Four Datasets. In the ground truth row, the percentage in 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 race.

Feature Extraction

The feature extraction process was standardized across all datasets to ensure comparability and reproducibility. We employed the RAPIDS platform [8], which supports a broad range of sensor data types and facilitates the integration of data streams from multiple devices. Our feature extraction focused on multiple behavioral dimensions, including:

- **Location:** Measured through the GPS data, providing insights into user mobility patterns, location variance, the radius of gyration, etc.
- **Phone Usage:** Monitored via app usage statistics and screen on/off events, helping infer social connectivity and daily routines.
- **Physical Activity and Sleep Patterns:** Collected through Fitbit devices, offering detailed metrics on physical movements, exercise routines, and sleep quality.

We incorporated multiple time windows when extracting the features, 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), the whole day, and the past two weeks.

Data Preparation

Each dataset underwent a cleaning and preparation phase to align them across different datasets, ensuring a consistent format for analysis. This process included normalization procedures to standardize feature scales across different datasets. After processing, the data is formatted as a time-series feature-vector matrix, paired with labels at certain timestamps. To standardize input shapes, we sliced the feature sequence into consistent backward four-week periods based on each label. We picked four weeks to cover previous depression detection models' feature calculations. Each label matches a feature matrix of identical shapes.

NOVEL GENERALIZABLE ALGORITHM: REORDER

The challenge in domain generalization is largely due to 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), but also from individual differences (i.e., each person behaves uniquely). Despite these differences, we observed similarities in behaviors across individuals. For example, although individuals have unique daily routines, these patterns lead to continuous behavior trajectories along the time domain. Such an observation motivates us to leverage behavior continuity and construct a self-supervised learning task to obtain generalizable feature representations.

The key idea behind **Reorder** is to augment the training process by incorporating a temporal reordering task, which compels the model to recognize and predict the correct sequence of observed behaviors. It shuffles the temporal order of the feature matrix, and trains a model to reconstruct the original sequence, jointly optimized with the main classification task.

Reorder achieves two tasks simultaneously: 1) it will learn to solve the main task (i.e., depression detection in our case); and 2) it will learn to capture the continuity of behavior trajectories, so that it can find the original temporal order of the time-series feature data before shuffling. Due to the prevalence of the continuous behavior trajectories based on human nature (analogous to the continuous edges and patterns in images [2]), solving the second task by learning such continuity could assist the model in extracting more generalizable representations of behavior trajectories across individuals. By focusing on the temporal dynamics of behavior, Reorder helps the model capture general characteristics that are invariant across different datasets and populations, thus enhancing the model's generalizability. Figure 3 illustrates its main concept compared to a vanilla deep learning model.

Implementation Details

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 $L_c(h(x|\theta_f,\theta_c), y)$, where x is the input matrix, and y is the classification label. The second task is the reordering task. Specifically, 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 still an overly large number, we predetermined a subset of P = 200 permutations by following the Hamming-distance-based method. 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 another classification task. Therefore, the loss function of the reordering task is $L_r(h(z|\theta_f,\theta_r), p)$, where z is the feature matrix x after the reordering, and p is the permutation index. Overall, the model can be trained via the following objective function:

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

where both L_c and 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 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.

OPEN-SOURCE BENCHMARK PLATFORM: GLOBEM

The *GLOBEM* platform (**G**eneralization of **LO**ngitudinal **BE**havior **M**odeling) offers a robust benchmarking tool for evaluating the generalizability of behavior models across diverse datasets. GLOBEM integrates our new Reorder algorithm along with other established and novel methodologies to create a comprehensive environment for development and evaluation. It is available at https://the-globem.github.io/.

GLOBEM is designed to support a wide array of algorithms, facilitating direct comparisons under standardized conditions. Other than Reorder, this platform includes nine traditional behavior modeling algorithms: Canzian *et al.* [1], Saeb *et al.* [6], Farhan *et al.* [4], Wahle *et al.* [9], Lu *et al.* [5], Wang *et al.* [11], Xu *et al.* - Interpretable [15], Xu *et al.* - Personalized [14], and Chikersal *et al.* [3]. It also incorporates 8 recent algorithms in domain generalization techniques. Other than the basic ERM (Empirical Risk Minimization), GLOBEM includes Mixup, IRM (Invariant Risk Minimization), DANN (Domain-Adversarial Neural Network), CSD (Common Specific Decomposition), MLDG (Meta-Learning for Domain Generalization), MASF (Model-Agnostic Learning of Semantic Features, and Siamese Network. More details of these algorithms can be found in [13].

An essential feature of GLOBEM is its open-source nature, which encourages collaboration and innovation in the community. Researchers and developers can access the full suite of datasets, algorithms, and other modules on GLOBEM. They can reuse 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.

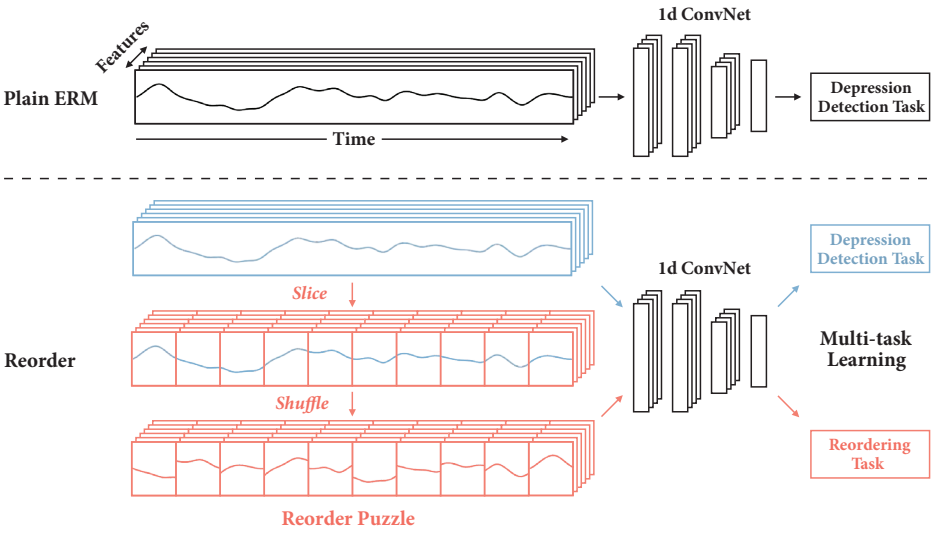


FIGURE 3. The Design of Reorder Compared to ERM. In addition to the main behavior modeling task, Reorder further introduces a secondary task of solving a reorder puzzle to force the model to learn the continuity of behavior trajectory.

EVALUATION

Dataset Analysis

In-depth analysis of four diverse datasets highlighted the complexities in behavioral features associated with depression. We first utilized linear mixed-effect models to quantify the strength and direction of relationships between individual features and depression. Key behavioral indicators such as sleep duration, phone usage, and physical activity showed statistically significant correlations with depression scores (see Figure 4.a). For instance, shorter sleep duration and higher frequency of phone usage were consistently associated with higher depression scores across all datasets.

The consistency of some features across datasets suggested universal behavioral markers of depression, while variations in other feature impacts underscored the influence of contextual factors specific to each dataset (see Figure 4.b). For example, the patterns of mobility (e.g., number of frequent locations visited) and physical activity (e.g., number of steps) varied across datasets, indicating potential cultural or environmental differences. Note that these datasets were all collected before COVID-19.

Single-Dataset Evaluation of Existing Algorithms

The performance analysis of existing depression detection models revealed significant discrepancies when applied to new datasets compared to their reported performance in their original studies. Models that achieved high accuracy in detecting depression in one dataset often failed to replicate this performance in another, indicating a lack of generalizability (Average $\Delta = 15.9 \pm 10.7\%$ for end-of-term depression prediction, and $\Delta = 22.6 \pm 8.5\%$ for weekly depression prediction), emphasizing the challenge of creating robust depression detection models that perform consistently across different populations and settings.

Cross-Dataset Evaluation

In cross-dataset evaluation, models were trained on three datasets and tested on the fourth. This setup simulated the challenge of deploying models in new environments where they had not been initially calibrated. Following are our primary observations.

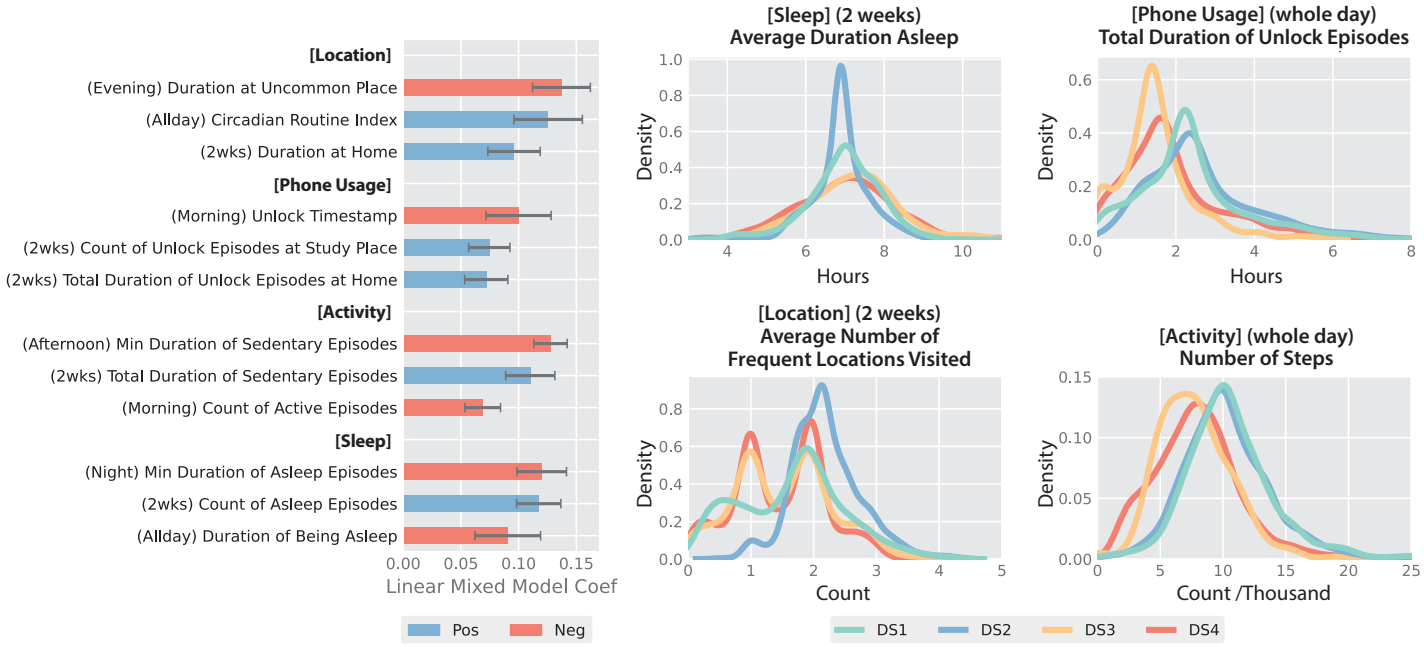
First, all nine depression detection models demonstrated worse performance

WE PRESENT GLOBEM [13], AN OPEN-SOURCE BENCHMARK PLATFORM THAT CONSOLIDATES A NUMBER OF ALGORITHMS TO FOSTER OPEN-SOURCE RESEARCH AND DEVELOPMENT IN THIS AREA. THIS PLATFORM ALLOWS FOR RIGOROUS EVALUATION ACROSS MULTIPLE DATASETS AND HEALTH PREDICTION TARGETS

than those in the single-dataset evaluation. The best model, *Chikersal et al.*, showed an average balanced accuracy of 52.0% and an ROC AUC of 54.1% in the cross-dataset evaluation, compared to 58.8% in the within-dataset evaluation.

Second, although modern ML techniques have been developed to deal with the challenge of feature shift across domains, these models did not work well on our datasets. Among the 15 models we investigated, *CSD - Person as Domain* and *ERM - 2D-CNN* achieved the highest ROC AUC (52.3%), similar to the results of traditional depression detection models (54.1% for *Chikersal et al.*). These evaluations illustrated that recent domain generalization methods do not work well on our datasets. Most of these methods were developed under the context of CV or NLP tasks, and their generalizability may be affected when applied to longitudinal behavior data.

Most importantly, of the 26 models evaluated, our newly proposed Reorder model achieved the highest ROC AUC of 57.5% and the highest balanced accuracy of 55.2%. As shown in Figure 5, Reorder stands out. 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



(a) Linear mixed model coefficiants of example features

(b) Distribution comparison of example features

FIGURE 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.

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.

Multiple Aspects of Generalization

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), and across years (i.e., different users within the same population). Moreover, in each institute, there was 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.

Figure 6 summarizes the key results. The cross-institute and cross-year evaluation results provide more insights into model generalizability. The model Reorder had 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 were clearly better than those of the other two setups, revealing that individual differences (no matter whether that is within or between populations) may play the most important role in the cross-dataset generalization challenge.

Even though Reorder shows promise in domain generalization, it is worth noting that our model still has room for improvement. The current performance is still far from being deployable in real-life scenarios, and we need more future research to improve model generalizability.

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 institutions, 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 a new method, Reorder, for better generalizability. Our evaluation of these models 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 in practical terms, with much room for improvement. Moreover, the comparison of multiple generalization tasks indicates that individual differences in behavior pose the main challenges for domain generalization in the longitudinal behavior modeling area. To assist future researchers in testing existing methods and developing new algorithms, we integrated all methods and open-sourced a benchmark platform named GLOBEM.

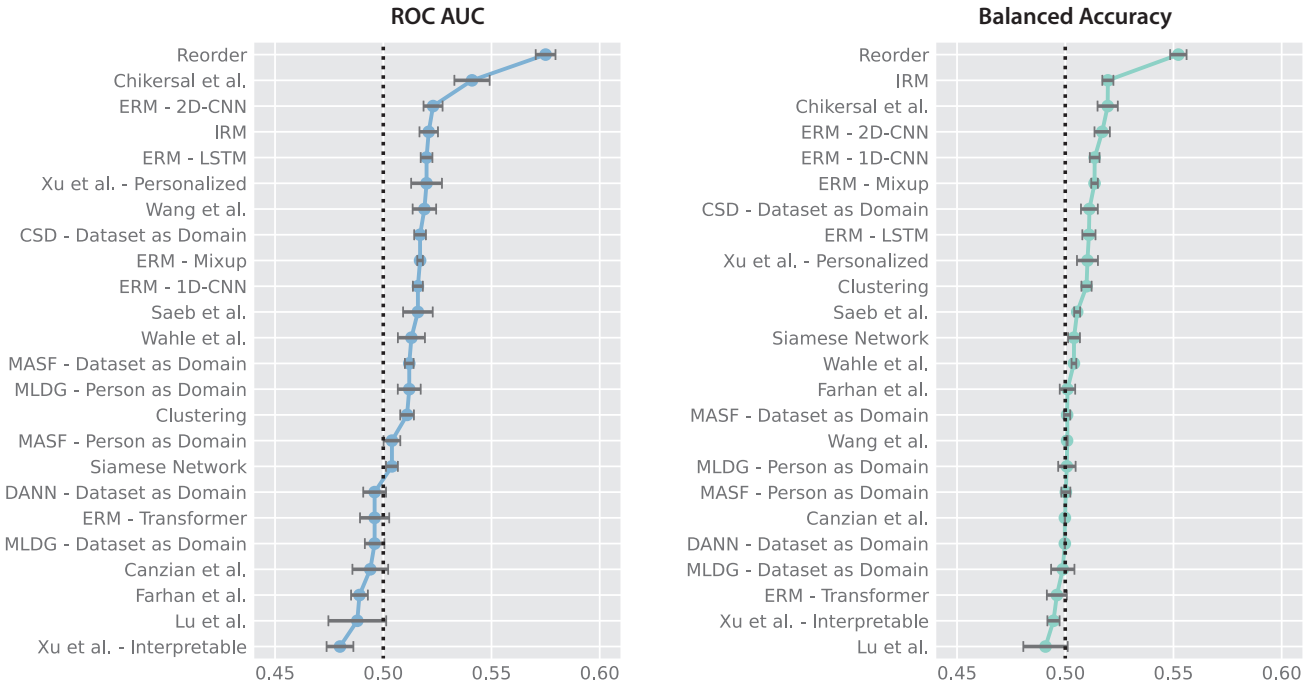


FIGURE 5. Model Performance of Predicting Bi-Weekly Depression Status across Datasets. The dashed line indicates a naive majority baseline. The same in Figure 6.

Call to Action

We encourage the research community to continue exploring these methodologies and to further develop the GLOBEM platform. By leveraging the collective expertise and resources of the community, we hope to accelerate the advancement of generalizable behavior models that are capable of supporting a wide range of applications, from mental health monitoring to personalized medicine.

This research sets the stage for a deeper investigation into the mechanisms that underpin effective behavior modeling across diverse datasets, toward the vision of ubiquitous sensing and analytics that support health and well-being on a global scale. ■

Xuhai (Orson) Xu is a postdoctoral associate at MIT. His research focuses on human-computer interaction, ubiquitous computing, applied machine learning, and health, specifically on developing intelligent behavior intervention systems to improve human well-being. He completed his PhD at the University of Washington.

Xin Liu is a Research Scientist at Google Consumer Health Research and a research affiliate at the University of Washington. His research interests lie in the intersection of machine learning, ubiquitous computing, and health. He completed his PhD at the University of Washington.

Han Zhang is a PhD student in Computer Science & Engineering at the University of Washington. Her research interests include human-computer interaction, human-centered machine learning, and the ethical aspects of artificial intelligence, specifically focusing on designing responsible tools to understand human behavior and support their well-being.

Weichen Wang is a Research Scientist at Meta. His research included projects on mobile sensing, wearable computing, and behavioral modeling to improve mental health outcomes. He completed his PhD at Dartmouth College.

Subigya Nepal is currently a PhD student in computer science at Dartmouth College. His research focuses on ubiquitous computing, passive sensing, and their applications in understanding and modeling human behavior.

Yasaman S. Sefidgar is a PhD candidate in Computer Science and Engineering at the University of Washington. Her research contributes theoretical frameworks, interaction techniques, and interface architectures that allow individuals to control data and AI systems and align these systems to their evolving needs.

Woosuk Seo is a PhD student at the University of Michigan School of Information, where he focuses on human-computer interaction, particularly in healthcare contexts. His research explores how technology can enhance the communication between children with cancer and their healthcare providers.

Kevin S. Kuehn is a T32-postdoctoral fellow in the Department of Medicine at the University of California, San Diego. His research focuses on suicidal thoughts and behaviors using intensive longitudinal methodologies. He obtained his PhD from the University of Washington.

Jeremy F. Huckins is a Vice President at Biocogniv Inc. His research is primarily focused on mental health and behavior, especially using mobile sensing technology to understand and address various psychological conditions. He completed his PhD at Dartmouth College.

Margaret E. Morris is an Associate Affiliate Professor at Information School, the University of Washington. She is a clinical psychologist who studies how technology can promote mental and physical health. She completed her PhD at the University of New Mexico, Albuquerque.

Paula S. Nurius is the Grace Beals-Ferguson Scholar and Professor at the University of Washington School of Social Work. Her research focuses on stress and trauma, especially among vulnerable and socially disadvantaged populations, aiming at early intervention and resilience building. She received her doctorate from the University of Michigan.

Eve A. Riskin is a Dean of Undergraduate Education and Professor of Electrical and Computer Engineering at Stevens Institute of Technology. Her research interests include image and video compression, human-computer interaction, and diversity in STEM. She completed her PhD at Stanford University.

Shwetak Patel is the Washington Research Foundation Entrepreneurship Endowed Professor at the University of Washington in Computer Science & Engineering and Electrical Engineering. His research interests include human-computer interaction, ubiquitous Computing, and sensor-enabled embedded systems. He completed his PhD at the Georgia Institute of Technology.

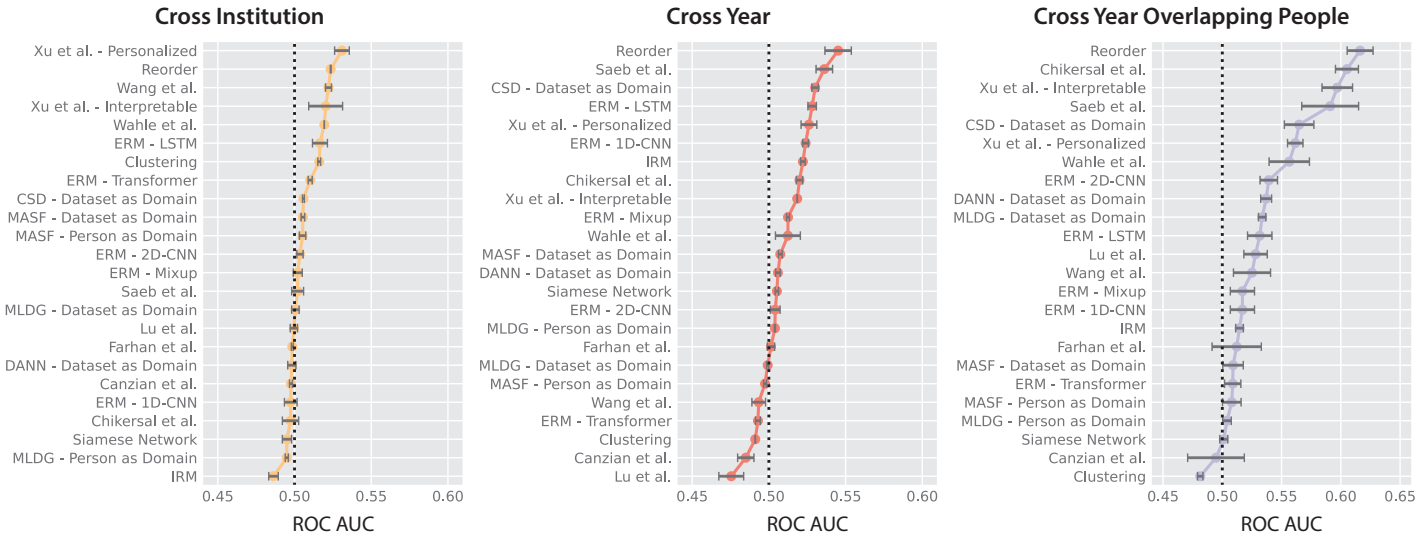


FIGURE 6. Model Performance of Predicting Bi-Weekly Depression Status across Institutions (left) and Years (middle, right). Models are tested on the datasets of one year/institution after being trained on the other year/institution.

Tim Althoff is an Assistant Professor in the Computer Science & Engineering at the University of Washington. His research focuses on better understanding and empowering people through data and computation. He completed his PhD at Stanford University.

Andrew T. Campbell is an Albert Bradley 1915 Third Century Professor in Computer Science Department at Dartmouth College. His research interests include using embedded sensors and machine learning on phones and wearables to infer human behavior with applications in health. He completed his PhD at the Lancaster University.

Anind K. Dey is a Professor and the Dean of the Information School at the University of Washington. His research focuses on context-aware computing, ubiquitous computing, and human-computer interaction. He completed his PhD at the Georgia Institute of Technology.

Jennifer Mankoff is the Richard E. Ladner Professor in the School of Computer Science & Engineering at the University of Washington. She directs the Center for Research and Education on Accessible Technology and Experiences. Her research focuses on assistive technologies, digital accessibility, sustainable technology, and fabrication. She completed her PhD at the Georgia Institute of Technology.

REFERENCES

[1] L. Canzian, M. 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*, 1293–1304. DOI:https://doi.org/10.1145/2750858.2805845.

[2] F.M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. 2019. Domain generalization by solving jigsaw puzzles. *arXiv*:1903.06864 [cs].

[3] P. Chikersal, A. Doryab, M. Tumminia, D.K. Villalba, J.M. Dutcher, X. Liu, S. Cohen, K.G. Creswell, J. Mankoff, J.D. Creswell, M. Goel, and A.K. Dey. 2021. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing. *ACM Transactions on Computer-Human Interaction*, 28, 1, 1–41. DOI:https://doi.org/10.1145/3422821.

[4] A.A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. 2016. Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. *IEEE Wireless Health (WH)*, 1–8.

[5] J. Lu, C. Bi, C. Shang, C. Yue, R. Morillo, S. Ware, J. Kamath, A. Bamis, A. Russell, and B. 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. DOI:https://doi.org/10.1145/3191753.

[6] S. Saeb, M. Zhang, C.J. Karr, S.M. Schueller, M.E. Corden, K.P. Kording, and D.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. DOI:https://doi.org/10.2196/jmir.4273.

[7] Y.S. Sefidgar, W. Seo, K.S. Kuehn, T. Althoff, A. Browning, E. Riskin, P.S. Nurius, A.K. Dey, and J. Mankoff. 2019. Passively-sensed behavioral correlates of discrimination events in college students. *Proceedings of the ACM on Human-Computer Interaction*. 3, CSCW (Nov. 2019), 1–29. DOI:https://doi.org/10.1145/3359216.

[8] J. Vega, M. Li, K. Aguillera, N. Goel, E. Joshi, K. Khandekar, K.C. Durica, A.R. Kunta, and C.A. Low. 2021. Reproducible analysis pipeline for data streams: Open-source software to process data collected with mobile devices. *Frontiers in Digital Health*. 3, 769823. DOI:https://doi.org/10.3389/fdgth.2021.769823.

[9] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, and S. Weidt. 2016. Mobile sensing and support for people with depression: A pilot trial in the wild. *JMIR mHealth and uHealth*. 4, 3 (2016), e111. DOI:https://doi.org/10.2196/mhealth.5960.

[10] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A.T. Campbell. 2014. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (New York, NY), 3–14.

[11] R. Wang, W. Wang, A. daSilva, J.F. Huckins, W.M. Kelley, T.F. Heatherton, and A.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, 1–26. DOI:https://doi.org/10.1145/3191775.

[12] M. Weiser, M. 1991. The computer for the 21st century. *Scientific American*, 265, 3 (1991), 94–105.

[13] X. Xu, et al. 2023. GLOBEM: Cross-dataset Generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 6, 4 (2023), 1–34. DOI:https://doi.org/10.1145/3569485.

[14] X. Xu, et al. 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, 1–27. DOI:https://doi.org/10.1145/3448107.

[15] X. Xu, P. Chikersal, A. Doryab, D.K. Villalba, J.M. Dutcher, M.J. Tumminia, T. Althoff, S. Cohen, K.G. Creswell, J.D. Creswell, J. Mankoff, and A.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, 1–33. DOI:https://doi.org/10.1145/3351274.