

Beyond Fine-Tuning: Participant-Specific Model Design for Mental Health Prediction

Erin Song
Northeastern University
Boston, MA, USA
song.erin@northeastern.edu

Lisa Marsch
Dartmouth College
Hanover, NH, USA
lisa.a.marsch@dartmouth.edu

Varun Mishra
Northeastern University
Boston, MA, USA
v.mishra@northeastern.edu

Abstract

Smartphones and wearables can capture behavioral and contextual states that show associations with various mental and affective health outcomes. Machine learning models have shown promise in detecting these outcomes, yet current approaches are commonly designed to generalize across users, often at the expense of capturing individual variability. Behavior varies across people, and is not always consistent. Over time, there may also be changes within the individual. Indeed, recent works have explored personalized modeling strategies. These approaches, however, often focus on limiting data to a specific participant, e.g., participant-specific training or fine-tuning, while keeping the model architecture and data-processing fixed. In this work, we explore how model design choices impact detection and prediction at the individual level. Specifically, we compare variations in class weighting, encoder design, look-back length, and number of hidden layers to assess what configurations best capture individual-level patterns. We evaluate these personalized models for a stress-prediction task, using data from a longitudinal study of 12 weeks with 58 participants undergoing opioid use disorder (OUD) treatment. Our preliminary results suggest that model configurations differ vastly across participants, highlighting the need to move beyond fine-tuning and participant-specific training to *participant specific model design*.

ACM Reference Format:

Erin Song, Lisa Marsch, and Varun Mishra. 2026. Beyond Fine-Tuning: Participant-Specific Model Design for Mental Health Prediction. In . ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Smartphones and wearable devices can passively measure several behavioral and physiological signals without being intrusive or overly burdensome, e.g., sleep, heart-rate, location, and physical activity [2–4]. Several prior works have studied the associations between these behavioral and physiological metrics obtained from passive sensing with subjective, self-reported outcomes from Ecological Momentary Assessments (EMAs) to measure mental health outcomes such as stress, craving, depression, and loneliness, and

have demonstrated feasibility in predicting these subjective outcomes using passive sensing data [2, 3, 8, 9, 13]. While promising, moving beyond feasibility and building systems and models that work reliably using passive sensing data alone is challenging. Passive sensing provides only a partial and often ambiguous view of a person’s context, while the outcomes of interest are subjective and self-reported. For instance, high physiological arousal could be due to excitement about a new project, exercising, giving a talk, or being stressed out for a paper submission [4]. Similarly, more time spent at home and less outdoor mobility could be due to onset of depressive symptoms, temporary change in work schedule, or winter weather [2, 3, 11]. A model that simply learns that high arousal indicates stress, or that low mobility indicates depression, may therefore capture patterns that are valid for some people in some contexts, but not for others. These relationships may also shift within the same person over time, as routines, contexts, and interpretations of internal states change. Thus, individual heterogeneity, contextual ambiguity, and temporal drift make mental health prediction inherently personalized, suggesting that models may need to be tailored to the individual rather than designed only to generalize across users.

Despite this individual and contextual variability, many current approaches use machine learning to generalize across users or across similarly grouped users [3, 9, 15]. These models often learn a single mapping between passive sensing features and self-reported outcomes, implicitly assuming that the same behavioral or physiological signals carry similar meaning across people. Prior work has recognized the limitations of these approaches to capture individual variability and have explored methods to personalize mental health predictions. These approaches typically personalize the data a model sees by crafting individual features, fine-tuning on a user’s own samples, or defining personal thresholds so that predictions are relative to the individual than the average [2, 22]. These approaches, however, often keep the broader model fixed, assuming that the same architecture, training procedure, data processing, temporal windows for feature calculation, etc., would work for everyone. This raises the question: Does one model configuration fit all individuals, or should the model design itself be personalized?

In this work, we examine how model design choices affect stress prediction at the individual level. We train and evaluate models on a per-participant basis, directly addressing the variability that pooled approaches obscure. Across participants we vary encoder design, training-window strategy, class weighting, and look-back length, and we compare what each configuration captures within and between individuals. We evaluate this approach while predicting stress in a 12-week longitudinal dataset of 58 patients undergoing opioid use disorder (OUD) treatment. Participants provide

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference’17, Washington, DC, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

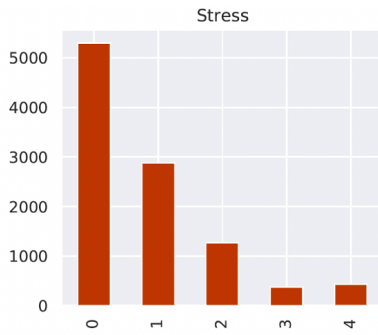


Figure 1: Overall stress response distribution for the entire study for all participants.

sensor data from smartphones and smart watches along with a maximum of three EMA responses throughout the day. Our preliminary analysis yields three key insights:

- Personalization work often keeps the model architecture, training procedures, and data processing choices fixed; however, the variables used to build and train the model are also unique to each person. The training techniques, not just the data and feature inputs, varies across people.
- People with similar stress patterns do not necessarily have the same best-performing model configuration.
- Participant-specific configuration selection improves performance compared with any single fixed configuration, especially by raising performance for participants who perform poorly under fixed model designs.

Together, these results suggest that choosing one configuration for the majority of participants can leave a subset of participants with substantially worse performance. We found that no single fixed configuration achieves the same overall AUROC distribution as the retrospective best-per-participant configuration. The goal of this *work-in-progress* paper is not to introduce a novel model or architecture that can improve stress prediction compared to current approaches, but to show that model design itself may be an important aspect of personalization. In particular, participant-specific model design may improve not only average performance, but also the lower end of the performance distribution for individuals who are poorly served by fixed configurations.

2 Related Work

Researchers have trained models with passive sensing from smartphones and wearables to detect a variety of mental health outcomes, e.g., stress, anxiety, depression, substance use, craving, etc. [1, 2, 5, 8–11, 13, 16] These works primarily rely on self-reports, often collected through EMAs, as the ground truth to train machine learning models [8, 17]. As the field has moved from controlled lab studies to free-living settings, this ground truth has become sparser and noisier, and reported accuracy has dropped accordingly [14]. Most of these models, however, are still trained to generalize across users. A single model learns one mapping from sensor features to an outcome and applies it to everyone in the study. This performs well on average, but results are often inconsistent across people,

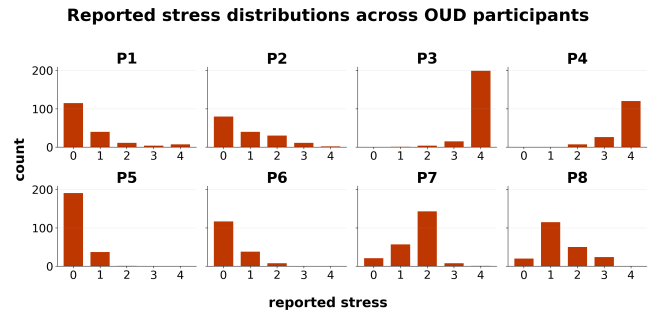


Figure 2: Stress response distribution of 8 participants.

and models trained on one population do not always reproduce on another [20]. These findings indicate that while passive sensing has potential, generalized models are not able to account for individual differences.

Prior work has recognized this limitation and responded by personalizing the data a model sees. Some approaches craft individual-level features or fine-tune a general model on a user’s own samples [2, 9, 22]. Others calculate personal thresholds, so that a label reflects a change from the individual’s own baseline rather than a population average [18]. A related line of work groups users before training, clustering people with similar traits or signal patterns and fitting a model to each group rather than to each person [3, 15]. These methods consistently improve on fully generalized models, which confirms that individual variability matters [2, 9]. Grouping, in particular, assumes that people whose data looks similar benefit from the same model, an assumption we return to in our findings.

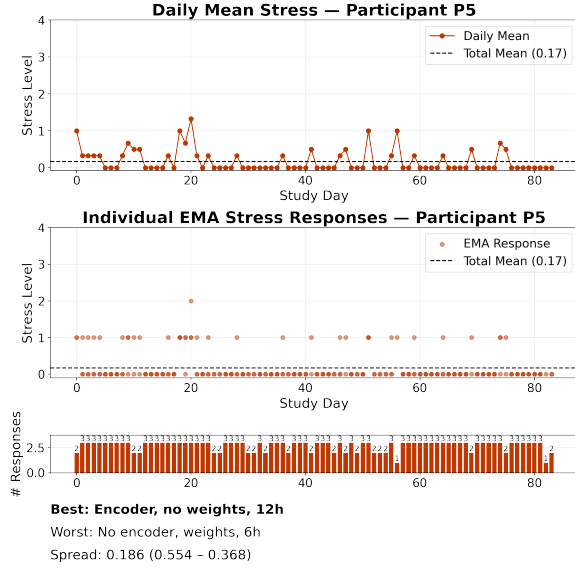
These solutions share a further assumption. The architecture, training window, label scheme, and other design choices are held fixed and presumed optimal for the given population. Personalization is applied to the input data, not to the model design itself. Some work does search over these design choices, through hyperparameter tuning or architecture search, but the search is run once and produces a single configuration applied to the larger group/study [2, 9, 11]. Hyperparameter tuning is not done on the individual level. Whether the best configuration is also specific to the individual remains largely unexamined. In this work, we treat the model design itself as something that may need to vary from person to person.

3 Methods

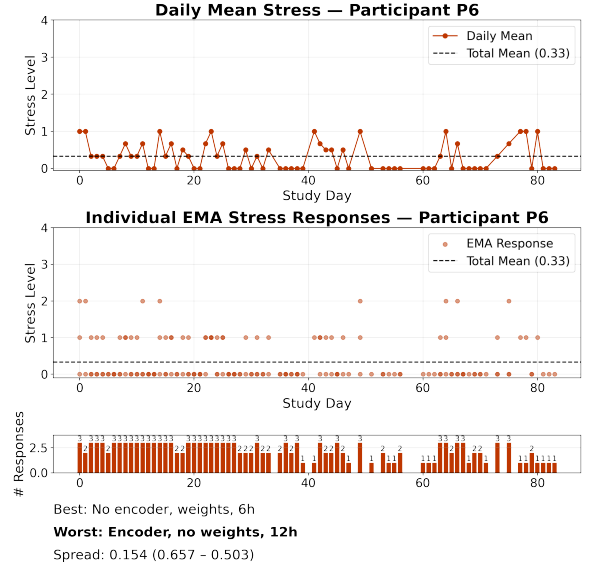
To test the effects of different architecture variables on personalized stress predictions, we apply multiple configurations to each participant and compare results. We observe both within and between patients and find differences in performance.

3.1 Dataset

For this work, we used the data collected by Campbell et al., from 58 patients with OUD over 12 weeks. In the study, the Campbell et al. provided participants with an Android phone and Garmin watch. They installed a study app that collected sensor data, e.g., GPS, activity, phone usage, ambient audio classifications, Bluetooth and Wi-Fi scans, step counts, sleep stages, and heart-rate. The study



(a) EMA stress responses for example participant 5.



(b) EMA stress responses for example participant 6.

Figure 3: Daily mean, EMA stress responses, and frequency of responses for two example participants. We observe that the best configuration for P5 is the worst for P6 and vice versa. The spread is the performance difference in the best and worst configuration.

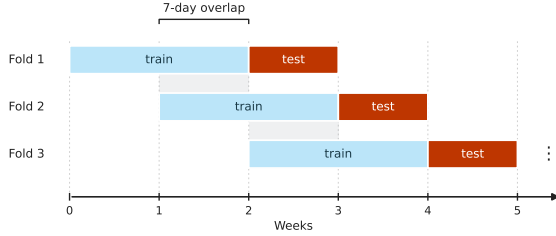


Figure 4: Rolling-window train/test split for each participant.

app also prompted each participant with three EMAs a day with multiple mental and behavioral health questions. For the scope of this work, we focused on the responses to stress. Each EMA prompt allowed the participants to provide their in-the-moment experience of stress on a five-point likert scale, ranging from “Not at all” (0) to “Extremely” (4).

3.2 Data Preprocessing

We preprocessed the sensing and EMA data by combining them into 15 minute buckets. This is due to the irregular frequencies of the sensors. Information such as heart rate was collected every 15 seconds, while steps was collected every 15 minutes. We combined the sensor data based on the smallest common denominator and created a fixed time series for each participant. Stress data was binarized as the overall response rate was heavily skewed left as shown by Figure 1. A response of “Not at all” (0) was considered “No stress” (0) and any response from 1 to 4 was considered “Some stress” (1). The data in each bucket was aggregated into simple features

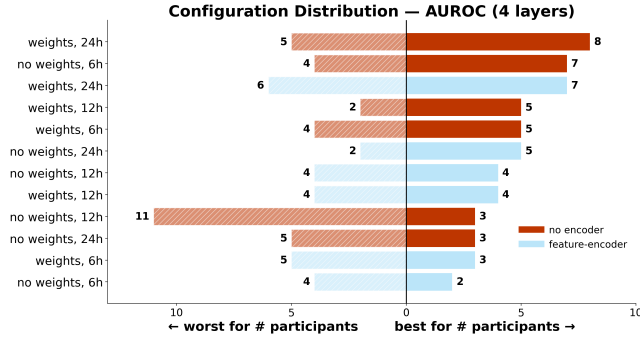
such as mean, median, mode, standard deviation, and other standard methods [6]. This way, we are able to capture information from all sensors for each bucket, rather than having multiple different data points that are not synced to any specific time.

3.3 Model Configurations

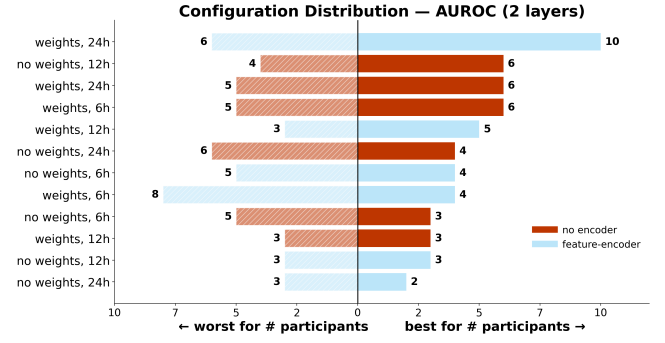
In this work, we use a standard Transformer model and adjust a few basic variables to test our hypothesis. We chose a transformer over a long short-term memory (LSTM) as they are better at capturing long-term dependencies and global temporal patterns [19]. We emphasize that the goal of this work is not to find the best performing model, but to highlight how the design and configuration of the model is important for personalization. We envision our findings would hold across different modeling approaches. For customization, we chose the following variables:

- Class weights (with or without)
- Feature encoder (with or without)
- Look-back period (6, 12, or 24 hours)
- Number of hidden layers

Class weighting adjusts the training loss so under-represented classes contribute more than their raw frequency would allow. It is commonly used when labels are imbalanced. Without weighting, a model can lower its loss by favoring the common response and rarely predicting the elevated one, which is the response we most want to detect. Weighting counteracts this by penalizing those missed cases more heavily. Given that we have binarized the data, some participants do not necessarily have a strong skew towards no stress (0) or some stress (1). Example distributions can be seen in Figure 2. The same setting that improves one participant’s model can degrade another, simply because their label distributions differ.



(a) Distribution of the best and worst configurations over all participants given two hidden layers in the transformer model.



(b) Distribution of the best and worst configurations over all participants given four hidden layers in the transformer model.

Figure 5: Configuration distribution showing the number of participants where the labeled configuration worked the best and worst for. The graphs are split by the number of hidden layers in the model for clarity.

Having an adjustable weights variable in the model configuration allows us to determine when having weights is reasonable or not.

The *feature encoder* is a small multi-layer perceptron applied to each 15-minute bucket. Each bucket is a vector of sensor features, and the encoder maps that vector into one learned representation before it enters the transformer. This allows the model to first learn within-bucket relationships among the behavioral features before modeling temporal relationships across buckets. Without the encoder, the transformer receives the raw features across time, and must learn both feature interactions and temporal dependencies simultaneously. With the encoder, the features in each bucket become one composed representation first, and attention then works over a sequence of distinct moments. This may matter more for some participants than others. If a participant’s stress is reflected in combinations of features within a moment, composing the bucket first may give the model a more useful unit to attend over. If their signal is carried by one or two features directly, the encoder may add little and, with sparse data, may instead learn noise. Like class weights, treating the feature encoder as an adjustable variable lets us examine when this additional representation-learning step is useful.

The *look-back period* sets how much history a tensor in the transformer model sees before each prediction, ranging from the most recent 6 hours to a full 24. A longer window gives the model more context, which matters when a state, such as stress, builds gradually. A shorter window focuses the model on recent signal, which matters when a state is driven by what happened in the last few hours. Which window is appropriate is specific to the person. Some participants may show stress that responds to immediate context, where a short look-back captures the relevant signal and a longer one only adds noise. Others may show stress that accumulates over a day, where a short window misses the build-up entirely. The window length also interacts with data availability: a longer look-back requires more continuous history per sample, so for a participant with frequent gaps it yields fewer usable samples even as it asks the model to learn from more. The best setting balances how far back a person’s relevant signal extends against how much complete data they actually have.

Furthermore, we also tested *two cases of hidden layers*. Looking at individual data means there is much less for the model to learn from, thus we first tested $n = 2$. We also tested $n = 4$ as the data itself is very complex and may have multiple correlations that cannot be captured at lower layer levels. This provides a comparison of changing model architecture, rather than only looking at varying training techniques.

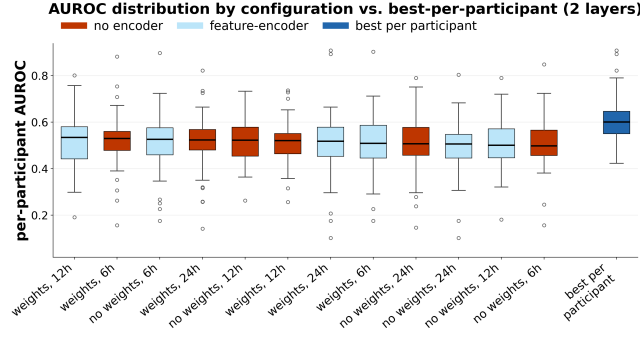
Beyond these four variables, there are also factors such as train-test split. Our design trains on two weeks of data and then tests on the third week with a one week overlap period. We show this structure in Figure 4. We did not test different configurations of the splits, however, it is possible that this is also a variable worth personalizing. By leaving the split static, we can first observe if the other common model configurations are unique to each participant. The best performing configurations and variables in our model are not necessarily universal.

4 Results

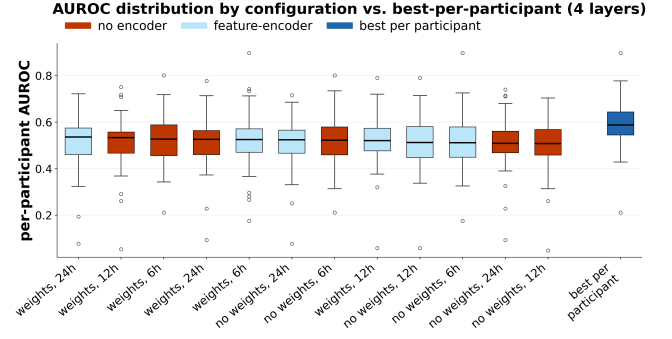
For each participant, we trained 24 configurations: two encoder settings, two class-weighting settings, three look-back lengths, and two layer settings, and identified the configuration that produced the best area under the curve (AUROC) for that person. We then examined how these best configurations were distributed across the 58 participants.

No single configuration worked best for a majority of participants. The most common configuration with a two layer transformer, used an encoder, class weighting, and a 24-hour look-back. This was optimal for only 10 of the 58 participants, or about 17% while also simultaneously being the worst for 10% of participants. The spread across all configurations can be seen in Figure 5a and Figure 5b. We observe a similar pattern for the four layer transformer. The variables also did not show a clear pattern in how often they mattered. Class weighting, feature encoding, and look-back length show no setting preferred across all participants. This pattern is what we would expect if the right configuration depends on the person rather than on the population.

Comparing individuals directly, we see that different configurations have varying results. For Participant 5 (see Figure 3a), the best



(a) Performance of each configuration given two hidden layers in the transformer model.



(b) Performance of each configuration given four hidden layers in the transformer model

Figure 6: Distribution of per-participant AUROC for all configurations as well as the distribution of the best configurations per participant. The graphs are split by the number of hidden layers in the model for clarity.

configuration used an encoder, no weights, and a 12-hour look-back (AUROC = 0.554), while the worst used no encoder, weights, and a 6-hour look-back (AUROC = 0.368). For Participant 6 (see Figure 3b), the best configuration used no encoder, weights, and a 6-hour look-back period (AUROC = 0.657), while the worst used an encoder, no weights, and a 12-hour look-back period (AUROC = 0.503). The configuration that was best for Participant 5 was the worst for Participant 6 and vice versa. The same model design that captured one participant’s patterns failed to capture the other’s, even though both participants have similarly shaped stress distributions.

Furthermore, the gap between the best and worst configuration was 0.186 for Participant 5 and 0.154 for Participant 6. These differences show that imposing one configuration across all participants can lead to suboptimal performance for several participants, especially those whose best configuration departs from the group default. Figure 6 shows this relationship directly. Each fixed configuration produces a wide spread of per-participant AUROC, with the lowest values often around 0.30. In contrast, the rightmost boxplot shows the AUROC each participant reaches under their own best-performing configuration, with the lower end of the distribution shifting upward relative to any single fixed configuration.

Importantly, we do not interpret these results as evidence that model design alone solves stress prediction for every participant. Some individuals remain difficult to predict, and our best per-participant results represent a retrospective upper bound rather than the performance a deployed system would achieve. Instead, our findings show that being deliberate about model design choices, even with relatively simple configurations, can improve overall performance and reduce the number of participants poorly served by a fixed model design. This suggests that model configuration is not incidental to personalization: the architecture, training choices, and temporal assumptions of a model may themselves need to vary across individuals. We expect this idea of participant-specific model design to remain relevant for more complex models, but further evaluation is needed to determine how to identify the right configuration for a new participant without exhaustively testing all options.

5 Discussion

Our results suggest that model configuration is a meaningful axis of personalization. Understanding which of these variables are worth tuning for a dataset can improve results for participants who may not perform as well under fixed conditions. Prior work has personalized the data a model sees, through individual features, fine-tuning, or personal thresholds. Our findings indicate that the model design itself may also need to vary from person to person. The same configuration that captured one participant’s stress patterns was the worst choice for another, and no single configuration worked best for more than about a quarter of our participants. If configuration is treated as fixed, this variability is lost, and any participant whose best design differs from the default is served by a model that underperforms for reasons beyond the quality or quantity of their data.

This may be specific to mental health. Stress is a subjective state, reported as it is perceived by the participant rather than measured directly, and it is expressed differently across people and across time within the same person. A model has to learn not only an individual’s signal but how that individual translates experience into a response. Tasks with objective, externally verifiable labels may not face this. In human activity recognition, walking produces a similar accelerometer signature across people, so a shared model design is a reasonable default [7, 12, 21]. Mental states have no such fixed signature. The added individuality we observe in model design may be a direct consequence of predicting something subjective, and we would not necessarily expect it for tasks where the label is grounded in the physical world.

Our work has several limitations: since we train one model per participant, the data available to each model is limited, which constrains the configurations we can fairly evaluate and likely affects which one appears best. Models for a participant with sparse data may favor a simpler configuration not because it suits them but because the alternatives had too little to learn from. There are also concerns regarding compute power. Running all permutations of the model variables to find the best fit structure for a participant will be taxing. In our results, we notice that there is a significant improvement for the lower performing participants. One potential

solution to compute limitations is to find the configuration that works best for the relative majority and then re-evaluate configurations for the worst performing participants. The best configuration for a specific participant may also change overtime due to factors such as temporal drift or changing context, meaning that each participant would have to have their own, continuously changing model structure. Another concern regarding mental health representation, is that the binarization of subjective feelings like stress, may not be the best solution. It may be necessary to also personalize the correlations between stress responses and other data. Finally, our findings are preliminary and drawn from a single OUD cohort. We do not claim that these specific configurations generalize; the claim is that the need to vary configuration does.

These limitations provide a guideline for future work. The central question is not only that configurations differ, but when and how to choose one for a new participant without testing all of them. A practical system cannot run an exhaustive per-person search, so future work should examine whether a participant's data characteristics, such as how skewed or sparse their responses are, can predict which configuration will serve them. There will also be differences in relevant variables depending on the base model used, as well as other variables that may be more relevant to predicting mental health outcomes. More broadly, these results suggest that participant-specific model design may be an important next step for personalized mental health prediction.

Acknowledgements

This research is partially supported by the National Institutes of Health, under award number NIDA P30DA029926 and the National Science Foundation, under award number IIS-2442593. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

- [1] Cynthia I. Campbell, Ching-Hua Chen, Sara R. Adams, Asma Asyied, Ninad R. Athale, Monique B. Does, Saeed Hassanpour, Emily Hichborn, Melanie Jackson-Morris, Nicholas C. Jacobson, Heather K. Jones, David Kotz, Chantal A. Lambert-Harris, Zhiguo Li, Bethany McLeman, Varun Mishra, Catherine Stanger, Geetha Subramaniam, Weiyl Wu, Christopher Zegers, and Lisa A. Marsch. Patient Engagement in a Multimodal Digital Phenotyping Study of Opioid Use Disorder. *Journal of Medical Internet Research*, 25(1):e45556, June 2023. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research.
- [2] 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. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection. *ACM Trans. Comput.-Hum. Interact.*, 28(1):3:1–3:41, January 2021.
- [3] Afsaneh Doryab, Daniella K. Villalba, Prerna Chikersal, Janine M. Dutcher, Michael Tumminia, Xinwen Liu, Sheldon Cohen, Kasey Creswell, Jennifer Mankoff, John D. Creswell, and Anind K. Dey. Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data. *JMIR mHealth and uHealth*, 7(7):e13209, July 2019.
- [4] Jessilyn Dunn, Varun Mishra, Md Mobashir Hasan Shandhi, Hayoung Jeong, Natasha Yamane, Yuna Watanabe, Bill Chen, and Matthew S. Goodwin. Building an open-source community to enhance autonomic nervous system signal analysis: DBDP-autonomic. *Frontiers in Digital Health*, 6, January 2025.
- [5] David H. Epstein, Matthew Tyburski, William J. Kowalczyk, Albert J. Burgess-Hull, Karran A. Phillips, Brenda L. Curtis, and Kenzie L. Preston. Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. *npj Digital Medicine*, 3(1):26, March 2020.
- [6] Simon Föll, Martin Maritsch, Federica Spinola, Varun Mishra, Filipe Barata, Tobias Kowatsch, Elgar Fleisch, and Felix Wortmann. FLIRT: A feature generation toolkit for wearable data. *Computer Methods and Programs in Biomedicine*, 212:106461, November 2021.
- [7] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Ir-fan Essa, Judy Hoffman, and Thomas Plötz. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers*, ISWC '20, pages 45–49, New York, NY, USA, September 2020. Association for Computing Machinery.
- [8] Michael V. Heinz, George D. Price, Avijit Singh, Sukanya Bhattacharya, Ching-Hua Chen, Asma Asyied, Monique B. Does, Saeed Hassanpour, Emily Hichborn, David Kotz, Chantal A. Lambert-Harris, Zhiguo Li, Bethany McLeman, Varun Mishra, Catherine Stanger, Geetha Subramaniam, Weiyl Wu, Cynthia I. Campbell, Lisa A. Marsch, and Nicholas C. Jacobson. A longitudinal observational study with ecological momentary assessment and deep learning to predict non-prescribed opioid use, treatment retention, and medication nonadherence among persons receiving medication treatment for opioid use disorder. *Journal of Substance Use and Addiction Treatment*, 173:209685, June 2025.
- [9] Boning Li and Akane Sano. Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2):49:1–49:26, June 2020.
- [10] Lisa A. Marsch, Ching-Hua Chen, Sara R. Adams, Asma Asyied, Monique B. Does, Saeed Hassanpour, Emily Hichborn, Melanie Jackson-Morris, Nicholas C. Jacobson, Heather K. Jones, David Kotz, Chantal A. Lambert-Harris, Zhiguo Li, Bethany McLeman, Varun Mishra, Catherine Stanger, Geetha Subramaniam, Weiyl Wu, and Cynthia I. Campbell. The Feasibility and Utility of Harnessing Digital Health to Understand Clinical Trajectories in Medication Treatment for Opioid Use Disorder: D-TECT Study Design and Methodological Considerations. *Frontiers in Psychiatry*, 13, April 2022.
- [11] Abhinav Mehrotra and Mirco Musolesi. Using Autoencoders to Automatically Extract Mobility Features for Predicting Depressive States. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3):127:1–127:20, September 2018.
- [12] Shenghuan Miao, Ling Chen, and Rong Hu. Spatial-Temporal Masked Autoencoder for Multi-Device Wearable Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(4):172:1–172:25, January 2024.
- [13] Varun Mishra, Tian Hao, Si Sun, Kimberly N. Walter, Marion J. Ball, Ching-Hua Chen, and Xinxin Zhu. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, pages 1708–1716, New York, NY, USA, October 2018. Association for Computing Machinery.
- [14] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. Continuous Detection of Physiological Stress with Commodity Hardware. *ACM Transactions on Computing for Healthcare*, 1(2):8:1–8:30, April 2020.
- [15] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. Evaluating the Reproducibility of Physiological Stress Detection Models. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(4):147, December 2020.
- [16] Anu Priya, Shruti Garg, and Neha Prerna Tigga. Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. *Procedia Computer Science*, 167:1258–1267, January 2020.
- [17] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, 2008.
- [18] Eden Tolosa, Isibor Kennedy Ihianle, Pedro Machado, Salisu Wada Yahaya, and Ahmad Lotfi. From Patterns to Deviations: Detecting Behavioural Drift for Mental Health Monitoring Using Smartphone and Wearable Data. *Electronics*, 15(4):885, January 2026.
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs.CL].
- [20] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve Riskin, Jennifer Mankoff, and Anind K. Dey. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization, March 2023. arXiv:2211.02733 [cs.LG].
- [21] Taeyoung Yeon, Vasco Xu, Henry Hoffmann, and Karan Ahuja. WatchHAR: Real-time On-device Human Activity Recognition System for Smartwatches. In *Proceedings of the 27th International Conference on Multimodal Interaction*, pages 387–394, October 2025. arXiv:2509.04736 [cs].
- [22] Han Yu and Akane Sano. Semi-Supervised Learning and Data Augmentation for Wearable-based Health Monitoring System in the Wild. December 2022.