Does Time Hurt Like People? Temporal Shift and Adaptation in Mobile Stress Sensing

Panyu Zhang
panyu@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Uichin Lee*
uclee@kaist.ac.kr
KAIST
Daejeon, Republic of Korea

Abstract

Mobile stress sensing models are often evaluated under cross-user generalization, where models trained on one set of users are tested on unseen users. However, real-world deployment is also longitudinal: models trained on earlier data may be applied later as users' behaviors, routines, and stressors change. In this preliminary study, we examine temporal shift in mobile stress sensing using the College Experience Study dataset. We compare cross-user 2-fold evaluation and temporal 50/50 chronological evaluation, while also reporting a random 2-fold split as an optimistic reference. We evaluate a baseline 1-D CNN model together with ten unsupervised domain adaptation methods. In this deployment-oriented comparison, chronological evaluation produced a larger reduction from the random reference than cross-user evaluation. User-as-domain variants consistently outperformed their two-domain counterparts, and all remained above the source-only baseline. Multi-branch adaptation provided further gains, particularly for DANN, although the best temporal AUROC remained below 0.60. These preliminary results suggest that temporal drift reveals a deployment challenge that cross-user evaluation alone may miss.

Public CES Dataset (original release): Kaggle | Our Code: GitHub

CCS Concepts

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; *Empirical studies in HCI*; • **Computing methodologies** → *Machine learning*.

Keywords

Stress Sensing, Health-Wellbeing, Temporal Shift, Mobile Sensing

ACM Reference Format:
Panyu Zhang and Uichin Lee. 2026. Does Time Hurt Like People? Temporal Shift and Adaptation in Mobile Stress Sensing. In *Proceedings of Adjunct Proceedings of the 2026 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2026 ACM International Symposium on Wearable Computing (UbiComp/ISWC '26).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*The corresponding author.

UbiComp/ISWC '26, Shanghai, China

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Mobile stress sensing aims to infer stress from passive mobile sensing signals collected in daily life [14]. Much of the robustness concern in this area has focused on *cross-user shift* [1, 4, 8, 18, 19]. Because users differ in behavior patterns and stress responses, models trained on one group of users may not transfer well to unseen users. In addition, the association between sensing features and self-reported stress or mental states can vary substantially across individuals [2]. For this reason, leave-user-out and user-disjoint evaluations have become common protocols for assessing whether mobile stress sensing models generalize beyond the users observed during training [19].

Yet deployment introduces another axis of generalization: *time*. A mobile human sensing model is often trained from past observations but used later, when the same user's behavior patterns or mental states may have changed [11]. Over time, both the sensing feature distribution and the behavior–stress relationship may drift, making later observations difficult to predict from earlier ones. Therefore, good performance on unseen users does not necessarily imply reliable performance in longitudinal deployment.

The College Experience Study (CES) provides a useful setting for studying this problem because it contains longitudinal mobile sensing and mental-health data from college students across multiple years and COVID-related periods [10]. Whereas prior CES work characterized longitudinal behavioral and mental-health changes, we frame its longitudinal structure as a predictive robustness problem. Using this dataset, we examine temporal shift as a deployment-oriented robustness problem for mobile stress sensing. We first compare how a baseline 1-D CNN model performs under random, cross-user, and temporal evaluation settings, focusing on whether the performance gap between random and temporal evaluation is comparable to the gap between random and cross-user evaluation. We then evaluate whether unsupervised domain adaptation (UDA) can improve performance under temporal shift.

UDA provides a natural starting point for studying temporal robustness because it is designed to address source-target distribution mismatch. Methods such as Domain-Adversarial Neural Networks (DANN) learn a feature extractor, label predictor, and domain classifier to reduce domain-specific information in the learned representation [3]. Prior mobile sensing adaptation work, such as M3BAT, further shows how adversarial adaptation can be applied in multimodal mobile sensing settings [9]. These methods motivate our temporal adaptation analysis, where earlier observations are treated as the source period and later observations as the target period.

In this paper, we conduct a preliminary empirical study of temporal shift and adaptation for mobile stress sensing using the CES

dataset. We report baseline 1-D CNN performance across random 2-fold, cross-user 2-fold, and temporal 50/50 chronological evaluation. We then evaluate ten UDA methods under the temporal setting. Our results show that chronological evaluation produces a larger reduction from the random reference than user-disjoint evaluation in this deployment-oriented comparison. User-as-domain alignment consistently improves methods with an explicit domain objective, while multi-branch adaptation provides additional gains, particularly for DANN.

Our study is guided by two research questions:

- **RQ1.** Will temporal shift affect model performance like cross-user shift in mobile stress sensing?
- **RQ2.** Do UDA methods provide gains in temporal shift settings?

Overall, this work argues that temporal shift should be treated as a first-class evaluation setting in longitudinal mobile stress sensing, rather than as a secondary robustness check after cross-user evaluation.

2 Dataset and Task

2.1 Dataset

We use the College Experience Study (CES), a longitudinal mobile sensing dataset collected from college students [10]. The original data contain 32,234 stress EMA observations from 161 users, spanning September 10, 2017 to June 25, 2022. During sequence construction, we excluded 370 observations whose 14-day windows contained fewer than four days with available sensing data. We further excluded one user with 238 observations for not satisfying the minimum requirement of 40 observations with both stress classes. The final analysis set therefore contains 31,626 prediction instances from 160 users.

Each instance is represented as a 14-day sequence of 17 daily sensing features, yielding an input of shape 14×17. Users contributed a median of 186 instances (IQR: 157–215; range: 101–440). In the fixed temporal evaluation, each user's chronologically ordered observations were divided into 15,806 source-period instances and 15,820 target-period instances. We verified that the observations of every user were strictly ordered by time before applying the chronological split.

A preliminary baseline screening across hourly, epoch-level, 7-day, and 14-day representations selected the 14-day daily input, which was fixed for all subsequent experiments.

2.2 Prediction Task

The prediction target is `stress_binary_personal`, a personalized binary stress label derived from the original self-reported stress score. Responses strictly above the applicable median are labeled as high stress. In all settings, each user's observations are divided into a calibration or training portion and a held-out evaluation portion. In the random setting, these portions are randomly sampled, whereas in the temporal setting the earlier observations form the source portion and the later observations form the target portion. In the cross-user setting, source and target users are disjoint for model training, but each target user retains a non-test calibration portion that is used only to estimate that user's stress threshold and feature-normalization statistics; the classifier is evaluated on the remaining held-out observations. Stress labels from the final test portion are never used to estimate thresholds or train the model. Each input instance is represented as a 14-day daily sequence of mobile sensing features with shape 14×17, and we report AUROC as a threshold-independent metric under label imbalance.

Under the fixed temporal split and the train-only label rule used in the experiments, 30.67% of source-period observations and 38.45% of target-period observations were labeled as high stress. This 7.78-percentage-point increase indicates a substantial temporal label-prevalence shift that may contribute to the observed performance degradation.

3 Evaluation Settings and Methods

3.1 Evaluation Settings

We evaluate three settings to distinguish an optimistic random reference, cross-user generalization, and longitudinal temporal generalization.

Random 2-fold split. We report a random 2-fold split as an optimistic reference. In this setting, samples are randomly divided into two folds, so the same users can appear in both training and testing, and temporal order is also mixed across the split. Therefore, this setting should not be interpreted as deployment performance. Instead, it provides an i.i.d.-style reference for the baseline model when both user identity and chronological ordering are mixed.

Cross-user 2-fold evaluation. In the cross-user setting, we use a user-disjoint 2-fold split. Training and testing users do not overlap, so this setting evaluates zero-shot generalization to unseen users. We use 2-fold evaluation to make the train/test proportion more comparable to the temporal 50/50 setting. This setting reflects a common mobile stress sensing scenario where a model trained on one group of users is evaluated on held-out users.

Temporal 50/50 chronological evaluation. In the temporal setting, each user's timeline is split chronologically. The first half of each user's timeline is used for training or adaptation, and the latter half is used for evaluation. Unlike the random split, this setting preserves temporal ordering. Unlike the cross-user setting, the evaluation data come from later observations in the same longitudinal cohort. We use this setting to examine whether models trained on earlier data remain predictive as users' routines, contexts, and stressors change over time.

Together, these settings allow us to compare the baseline model under three different assumptions: random sample mixing, unseen-user generalization, and future-time generalization. In particular, we compare the performance gap between random and temporal evaluation with the gap between random and cross-user evaluation, rather than treating the random split as a deployment setting.

The baseline comparison in Table 1 uses two folds with one training run per fold and fixed seed 42. The reported mean and sample standard deviation therefore summarize two fold-level pooled AUROCs.

The UDA comparison in Table 2 uses one deterministic temporal 50/50 split and five paired repetitions with seeds 42–46. For each repetition, 20% of the labeled source observations are selected as a stratified source-validation set. Each method independently

performs 30 Optuna trials. Hyperparameter optimization, early stopping, and checkpoint selection use source-validation AUROC, with a patience of 15 epochs and a maximum of 50 epochs. UDA training uses labeled source observations and unlabeled target features. Target labels are accessed only after checkpoint selection for final AUROC evaluation.

3.2 Baseline 1-D CNN Model

All experiments use a compact 1-D CNN sequence model as the shared backbone. Each input instance has shape $(B, T = 14, F = 17)$, where B is the batch size, T is the number of daily time steps, and F is the number of daily features. The input is transposed to $(B, 17, 14)$ before being passed to the temporal convolutional encoder.

The encoder consists of one or two Conv1d blocks with batch normalization, ReLU activation, and dropout, followed by adaptive average pooling. The pooled representation is passed through a single linear projection to obtain a 128-dimensional embedding, followed by a single linear classifier for binary stress prediction. The number of convolutional blocks, hidden dimension, dropout, learning rate, and weight decay are selected using source-validation HPO. We refer to this model as the **baseline 1-D CNN**.

3.3 UDA Methods

After comparing the baseline 1-D CNN across the three evaluation settings, we focus on the temporal setting to test whether UDA methods can improve longitudinal generalization. In this setting, the first chronological half of each user's data is treated as the source period, and the latter half is treated as the target period. Target labels are used only for evaluation.

We evaluate DANN [3], CDAN [7], DeepCORAL [15], ADDA [16], MCC [5], MCD [13], SHOT [6], CBST [20], DANN-M3BAT, and CDAN-M3BAT [9]. For methods with an explicit domain objective, we treat individual users as domains. DANN and CDAN use multiclass user discriminators, with CDAN operating on class-conditioned features; DeepCORAL aligns per-user feature statistics; and ADDA uses multiclass user discrimination with uniform-confusion training. The M3BAT variants additionally use high-, medium-, and low-shift feature branches weighted by source–target Cohen's d. MCC, MCD, SHOT, and CBST remain unchanged because their objectives contain no domain variable.

4 Results

4.1 RQ1: Temporal Shift Substantially Affects Baseline Performance

Table 1 summarizes the baseline 1-D CNN performance across the three evaluation settings using leakage-free personalized stress labels. The random 2-fold split achieves the highest AUROC of 0.6606 ± 0.0027. However, this setting mixes both user identity and temporal order across train and test, so it should be interpreted as an optimistic reference rather than deployment performance.

Compared with the random 2-fold split, the baseline AUROC drops to 0.6081 ± 0.0044 under cross-user 2-fold evaluation, corresponding to a decrease of 5.25 percentage points. Under temporal 50/50 chronological evaluation, the AUROC further drops to 0.5801 ± 0.0051, corresponding to a decrease of 8.05 percentage points from the random reference. Thus, the performance gap between random and temporal evaluation is larger than the gap between random and cross-user evaluation. This suggests that temporal shift can introduce degradation comparable to, or larger than, the standard unseen-user generalization challenge.

Table 1: Baseline 1-D CNN performance across evaluation settings. Means and sample standard deviations summarize two fold-level pooled AUROCs with one run per fold. With only two observations, the dispersion estimates are descriptive. The random split is an optimistic reference rather than a deployment setting.

Setting	Baseline AUROC	Drop from Random
Random 2-fold split	0.6606 ± 0.0027	–
Cross-user 2-fold	0.6081 ± 0.0044	−5.25 pp
Temporal 50/50	0.5801 ± 0.0051	−8.05 pp

Importantly, the random split is not a no-shift condition and does not provide a causal decomposition of user and temporal effects. It only serves as an optimistic reference. Under this interpretation, temporal evaluation reveals a substantial longitudinal deployment challenge that would be missed by reporting only random or cross-user performance.

4.2 RQ2: UDA Provides Limited Gains under Temporal Shift

Table 2 reports five paired repetitions on one fixed temporal split, with hyperparameter optimization, early stopping, and checkpoint selection based exclusively on source-validation AUROC. DANN-M3BAT achieves the highest mean AUROC of 0.5921 ± 0.0060, followed by CDAN-M3BAT at 0.5897 ± 0.0056. These correspond to numerical gains of 1.20 and 0.96 percentage points over the source-only baseline of 0.5801 ± 0.0045.

All six methods using users as domains remain above the baseline. Compared with their binary source–target reference implementations, ADDA improves from 0.5648 to 0.5846 (+1.98 pp), CDAN from 0.5800 to 0.5871 (+0.71 pp), DeepCORAL from 0.5807 to 0.5838 (+0.31 pp), and DANN from 0.5822 to 0.5835 (+0.13 pp). In contrast, none of the methods without an explicit domain term provides a meaningful gain over the baseline.

The ablation further shows that the two adversarial families benefit differently. For CDAN, most of the improvement comes from using individual users as domains (+0.71 pp), while the multi-branch encoder adds another 0.26 pp. For DANN, the user-as-domain formulation adds only 0.13 pp, whereas multi-branch processing contributes a further 0.86 pp.

Because all repetitions share the same temporal boundary, target observations, and participant cohort, the reported variability reflects source-validation splitting, hyperparameter optimization, and initialization rather than uncertainty across temporal periods or populations. Overall, user-as-domain alignment consistently improves methods with explicit domain objectives, but the remaining longitudinal performance gap is still substantial.

Table 2: Temporal 50/50 AUROC over five paired repetitions on one fixed chronological split. Methods with an explicit domain objective use individual users as domains; M3BAT variants additionally use multi-branch encoders. Δ is relative to the source-only baseline.

Method	Temporal AUROC	Δ
DANN-M3BAT	**0.5921 ± 0.0060**	+1.20 pp
CDAN-M3BAT	0.5897 ± 0.0056	+0.96 pp
CDAN	0.5871 ± 0.0040	+0.70 pp
ADDA	0.5846 ± 0.0064	+0.45 pp
DeepCORAL	0.5838 ± 0.0045	+0.37 pp
DANN	0.5835 ± 0.0091	+0.34 pp
SHOT	0.5806 ± 0.0048	+0.05 pp
Baseline 1-D CNN	0.5801 ± 0.0045	–
MCC	0.5782 ± 0.0043	−0.19 pp
MCD	0.5781 ± 0.0041	−0.20 pp
CBST	0.5710 ± 0.0121	−0.91 pp

5 Discussion

5.1 Temporal Shift as a Deployment Challenge

Our results suggest that temporal shift is not a secondary detail in mobile stress sensing. Relative to the random 2-fold reference, the baseline 1-D CNN shows a larger performance drop under temporal 50/50 evaluation than under cross-user 2-fold evaluation. This does not mean that the random split is a no-shift condition, nor does it causally decompose user shift and temporal shift. Rather, it shows that when temporal order is respected, longitudinal deployment can reveal a degradation that is comparable to, or larger than, the degradation observed under standard unseen-user evaluation. Therefore, cross-user evaluation alone may miss an important failure mode of mobile stress sensing systems.

5.2 User-as-Domain Alignment under Temporal Shift

The ablation suggests that binary source–target alignment is poorly matched to the per-user temporal setting. Because the same users appear in both periods, binary alignment pools substantial inter-user heterogeneity into only two temporal domains. Treating each participant as a domain instead directly encourages user-invariant representations, and every method supporting this formulation improves over its binary source–target counterpart.

ADDA provides the clearest example: its binary-domain reference performs worst overall, whereas its user-as-domain formulation reaches 0.5846, above the source-only baseline. This recovery suggests that inter-user variability provides a more useful alignment signal than the pooled earlier-versus-later distinction. The M3BAT ablation further shows that user-as-domain alignment and multi-branch processing provide complementary benefits, although their relative importance differs between DANN and CDAN.

Nevertheless, the best temporal AUROC remains below 0.60. User-as-domain alignment therefore closes only part of the longitudinal gap, and practical deployment would still require recalibration, sparse user-label collection, uncertainty-aware decision support, or selective model updating.

5.3 Limitations

This work has several limitations. First, it is a preliminary poster study based on the CES dataset and one binary stress label. We focus on CES because it provides a rare same-user, multi-year longitudinal mobile sensing setting, which is well suited for studying temporal shift in deployment. However, results from one longitudinal cohort should still be interpreted as an initial case study and validated on additional long-term mobile sensing datasets when such datasets become available.

Second, the random 2-fold, cross-user 2-fold, and temporal 50/50 settings have different deployment meanings: the random split mixes both user identity and temporal order, while the cross-user and temporal settings capture different practical generalization scenarios. We therefore interpret the results as a deployment-oriented comparison, not as a controlled causal decomposition of user shift and temporal shift.

Third, 2-fold evaluation provides limited variance estimation and is not sufficient for strong statistical claims. Finally, the current analysis does not isolate whether the observed degradation is primarily associated with feature shift, label-prevalence shift, missingness, or changes in the behavior–stress relationship. Although CES spans COVID-related periods, the present split does not isolate a causal COVID effect.

5.4 Future Work: Temporal Update Policies

Our results motivate moving from offline temporal benchmarking to deployment-time update policies. One option is fixed-interval adaptation, where models are periodically refreshed using newly collected data. For example, Time2Stop updates adaptive models at scheduled intervals to account for evolving user behavior [11]. Mobile stress models could similarly be updated after fixed time windows or a specified number of new ESM responses.

Alternatively, OOD-triggered adaptation could update the model only when feature, embedding, output, or uncertainty distributions indicate meaningful shift [12]. Detected shifts could trigger test-time or source-free adaptation, such as TENT [17] or SHOT [6], or sparse user-label collection. A hybrid policy could combine periodic refreshes with earlier OOD-triggered updates and decide when to retain, adapt, or recalibrate the deployed model.

6 Conclusion

This preliminary study examined temporal shift and adaptation in mobile stress sensing using the College Experience Study dataset. Our baseline results show that temporal 50/50 evaluation produces a larger performance drop from the random 2-fold reference than cross-user 2-fold evaluation, suggesting that longitudinal drift can be at least as challenging as unseen-user generalization in this focused comparison.

Methods supporting user-as-domain alignment consistently outperform their binary source–target counterparts, while multi-branch adaptation provides further gains, particularly for DANN. DANN-M3BAT achieves the strongest temporal result, but the absolute AUROC remains below 0.60. These findings suggest that inter-user variability is an important adaptation signal under longitudinal deployment, while temporal shift should remain a first-class evaluation setting in mobile stress sensing.

References

[1] Daniel A. Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C. Mohr, Anna R. Van Meter, Julia C. Tartaglia, Nicholas C. Jacobson, Fei Wang, Deborah Estrin, and Tanzeem Choudhury. 2024. Beyond Detection: Towards Actionable Sensing Research in Clinical Mental Healthcare. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 160 (Nov. 2024), 33 pages. doi:10.1145/3699755

[2] Isabela Albuquerque, João Monteiro, Olivier Rosanne, and Tiago H. Falk. 2022. Estimating distribution shifts for predicting cross-subject generalization in electroencephalography-based mental workload assessment. *Frontiers in Artificial Intelligence* 5, Article 992732 (Oct. 2022). doi:10.3389/frai.2022.992732

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* 17, 59 (2016), 1–35.

[4] Yunjo Han, Panyu Zhang, Minseo Park, and Uichin Lee. 2024. Systematic Evaluation of Personalized Deep Learning Models for Affect Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 206 (Nov. 2024), 35 pages. doi:10.1145/3699724

[5] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. 2020. Minimum Class Confusion for Versatile Domain Adaptation. In *Computer Vision – ECCV 2020 (Lecture Notes in Computer Science, Vol. 12354)*. Springer, 464–480.

[6] Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 6028–6039.

[7] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. 2018. Conditional Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems*. 1647–1657.

[8] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 176 (Jan. 2023), 32 pages. doi:10.1145/3569483

[9] Lakmal Meegahapola, Hamza Hassoune, and Daniel Gatica-Perez. 2024. M3BAT: Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 46 (May 2024), 30 pages. doi:10.1145/3659591

[10] Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F. Huckins, Courtney Rogers, Meghan L. Meyer, and Andrew T. Campbell. 2024. Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 38 (March 2024), 37 pages. doi:10.1145/3643501

[11] Adiba Orzikulova, Han Xiao, Zhipeng Li, Yukang Yan, Yuntao Wang, Yuanchun Shi, Marzyeh Ghassemi, Sung-Ju Lee, Anind K Dey, and Xuhai Xu. 2024. Time2Stop: Adaptive and Explainable Human-AI Loop for Smartphone Overuse Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 250, 20 pages. doi:10.1145/3613904.3642747

[12] Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. 2019. *Failing loudly: an empirical study of methods for detecting dataset shift.* Curran Associates Inc., Red Hook, NY, USA.

[13] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3723–3732.

[14] Akane Sano and Rosalind W. Picard. 2013. Stress Recognition Using Wearable Sensors and Mobile Phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 671–676. doi:10.1109/ACII.2013.117

[15] Baochen Sun and Kate Saenko. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Computer Vision – ECCV 2016 Workshops (Lecture Notes in Computer Science, Vol. 9915)*. Springer, 443–450.

[16] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial Discriminative Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7167–7176.

[17] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:232278031

[18] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, 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. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 190 (Jan. 2023), 34 pages. doi:10.1145/3569485

[19] Panyu Zhang, Gyuwon Jung, Jumabek Alikhanov, Uzair Ahmed, and Uichin Lee. 2024. A Reproducible Stress Prediction Pipeline with Mobile Sensor Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 143 (Sept. 2024), 35 pages. doi:10.1145/3678578

[20] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. 2018. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 289–305.