Training-Free Few-Shot Personalization for Mobile Mood Inference with TabPFN

Anonymous Author(s)

Abstract

Mobile sensing models often perform poorly for users and geographic populations not represented during model development. Prior work on DiversityOne showed that partial personalization with a Random Forest Hybrid Model improves mood inference by incorporating labeled examples from target users. However, this Hybrid Model must fit a new task-specific classifier whenever the target user, source population, or personalization budget changes. We investigate whether TabPFN can instead use the same labeled target-user examples through in-context learning, without task-specific parameter updates. We reconstruct the Hybrid Model on the DiversityOne release and compare it with TabPFN under matched experimental conditions: identical preprocessing, source-population data, labeled target-user personalization data, and fixed held-out target-user test data. We evaluate two-class and three-class valence prediction across six geographic configurations: country-specific, continent-specific, Country-Agnostic I, Country-Agnostic II, natural multi-country pooling, and balanced multi-country pooling. To examine the effect of labeled target-user support data, we report results separately at five personalization budgets: 10%, 20%, 30%, 40%, and 50%. Across these settings, TabPFN generally matches or improves over the Random Forest Hybrid Model, with the most consistent gains in Country-Agnostic I and balanced multi-country pooling. However, its advantage is not universal: performance depends on task formulation and pooling strategy, and the Hybrid Model remains stronger in some pooled three-class settings. These findings position tabular in-context learning as a strong training-free personalization baseline for mobile mood inference.

Resources: DiversityOne dataset [Project Website] | Our code [GitHub]

CCS Concepts

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; *Ubiquitous and mobile computing systems and tools*; • **Computing methodologies** → *Machine learning*; • **Applied computing** → Health informatics.

Keywords

Mobile Sensing, Mood Inference, Personalization, Tabular In-Context Learning, Geographic Generalization

Conference acronym 'XX, Woodstock, NY

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

ACM Reference Format:
Anonymous Author(s). 2018. Training-Free Few-Shot Personalization for Mobile Mood Inference with TabPFN . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Passive smartphone sensing can support continuous inference of mood and other affective states [9]. However, real-world deployment frequently requires prediction for users who were not represented during model development. New users and populations may differ in mobility routines, communication behavior, device-use patterns, daily schedules, cultural and geographic context, and the relationships between observable behavior and mood [1, 7, 8, 10]. These differences make population-level mobile sensing models difficult to transfer directly to unseen users and countries [2, 4, 12].

Prior DiversityOne research established that population-level models perform poorly for unseen users and countries, while adding labeled target-user data through a Random Forest Hybrid Model substantially improves mood inference [7]. The conventional Hybrid Model combines source-population data, labeled examples from the target user, and a newly fitted Random Forest classifier [7]. Because labeled target-user examples are directly included in training, the Hybrid Model requires retraining whenever the target user, source population, or amount of labeled target-user personalization data changes. This motivates a training-free alternative for partial personalization.

Recent work has explored large language models (LLMs) for digital mental health, wearable sensing, and affective computing. Mental-LLM shows that pretrained LLMs can support mental health prediction from online text data, but does not address passive smartphone sensing [11]. Health-LLM extends LLM-based prediction to wearable and physiological data, but relies on prompting or fine-tuning [6]. Other recent work suggests that LLMs can also predict affective states from smartphone sensor features [13]. However, these approaches typically convert structured sensing features into textual prompts or adapt language models.

We study a complementary direction: tabular in-context learning for mobile sensing personalization. TabPFN is pretrained for tabular prediction and conditions directly on labeled examples during inference [5]. Unlike the Random Forest Hybrid Model, TabPFN does not fit a new classifier for each personalization setting. Unlike LLM-based approaches, it operates directly on structured sensing-feature rows rather than textualized prompts. To our knowledge, tabular in-context learning has not previously been studied for personalization in mobile affective and mental sensing. The open question is whether TabPFN can extract more predictive value from limited labeled target-user examples than conventional classifier fitting.

We address two research questions:

- **RQ1:** How does TabPFN compare with the conventional Random Forest Hybrid Model across personalization budgets and geographic modeling settings?
- **RQ2:** Can TabPFN personalize to a new user using only that user's labeled data, without any source-population data?

For RQ1, we compare TabPFN and the Hybrid Model under matched experimental conditions: identical source-population data, labeled target-user data, fixed held-out target-user test data, feature representations, and preprocessing. We evaluate five personalization budgets: 10%, 20%, 30%, 40%, and 50%. We further evaluate six geographic configurations: country-specific, continent-specific, Country-Agnostic I, Country-Agnostic II, natural multi-country pooling, and balanced multi-country pooling [7]. For RQ2, we compare target-only TabPFN with a target-only Random Forest across the same five support budgets and multiple geographic pooling settings, using identical target-user support and test splits and excluding all source-population data.

This paper makes three contributions:

- To our knowledge, we present the first study of tabular in-context learning for mobile affective and mental sensing.
- We conduct a paired fixed-test evaluation against the established Random Forest Hybrid Model across five personalization budgets, six geographic configurations, and two valence-classification tasks.
- We evaluate target-only personalization across support budgets and geographic pooling settings to determine whether TabPFN can construct personalized predictors without source-population data.

2 Methodology

2.1 Dataset and Prediction Tasks

We use the released DiversityOne smartphone sensing dataset [1], which contains passive sensing and time-diary self-report data collected from participants in eight countries: China, Denmark, India, Italy, Mexico, Mongolia, Paraguay, and the United Kingdom. Our prediction target is mood valence derived from time-diary self-reports.

We evaluate two valence-classification tasks. For the two-class task, valence values 1, 2, 3 are mapped to the positive class and values 4 and 5 are mapped to the negative class. For the three-class task, values 1 and 2 are mapped to positive, value 3 is mapped to neutral, and values 4 and 5 are mapped to negative. The primary metric is macro AUROC. For the three-class task, AUROC is computed using a one-versus-rest formulation. Unless otherwise stated, results are averaged over 10 repeated user-level random holdouts. In each repeat, approximately 20% of eligible users are randomly selected as unseen target users, and the remaining users form the source population. We report mean macro AUROC with standard error across valid repeats.

2.2 Feature Reconstruction and Filtering

We reconstruct features from the released raw sensing streams. The final feature representation contains approximately 102 features when the required sensors are available [7]. The reconstructed modalities cover location, connectivity, proximity, notifications, activity, screen/touch interaction, user presence, episode duration, and app categories. Following the original DiversityOne feature design, features are aggregated within ±5-minute windows around each time-diary report [7].

Our reconstruction differs from the original generalization and personalization study in filtering and preprocessing. The original study reported 678 participants and 329,974 self-reports across eight countries [7]. Our reconstruction first retains valid mood answers, yielding 621 users and 279,585 self-reports; requiring observed sensor data in the ±5-minute window leaves 221,197 valid sensing windows. After task-specific label construction and minimum-report filtering, the final evaluated pool contains 526 users and 221,175 self-reports for both two-class task and three-class tasks. Missing values are median-imputed using only training or context rows, without target-test information. We therefore treat our experiments as a public-release reconstruction rather than an exact replication of the original feature table.

2.3 Reconstructed Hybrid Model Baseline

We reconstruct the Random Forest Hybrid Model (HM) from DiversityOne as the reference personalization baseline [7]. HM uses both source-user records and labeled target-support records. When target-support records are added, we remove an equal number of randomly chosen source records (sampled with a per-repeat seed), so that the total number of training or context records stays fixed at the size of that repeat's source pool. The total is therefore set by the setting and the repeat rather than by the support budget, ranging from approximately 2.6K records in the smallest country-specific configuration to approximately 217.8K records in Country-Agnostic II. This keeps the budget sweep from varying context volume, while the fixed test set isolates the effect of added target-user data. Both models receive exactly the same records under this rule.[1] The Random Forest uses 100 trees.

Our reconstruction reproduces the original personalization effect, although not every absolute AUROC value. Under the filtered pipeline used in our main experiments, aggregate PLM AUROC is 0.531/0.520 for two-/three-class prediction, close to the original 0.510/0.518, while HM reaches 0.699/0.717. When the additional filtering is removed, HM increases to 0.862/0.858, substantially narrowing the discrepancy with the original results. Thus, absolute HM performance is sensitive to reconstruction and filtering choices, while the central personalization effect is consistently reproduced.

2.4 Fixed-Test Target-Support Protocol

We use 10 repeated user-level random holdouts. In each repeat, approximately 20% of eligible users are selected as unseen target users, while the remaining users form the source population. For each target user, a fixed stratified 50% of eligible records is reserved as the target-test set, and the remaining 50% forms the support pool.

We sample target-support records at 10%, 20%, 30%, 40%, and 50% of each user's eligible records; the 50% budget uses the full support pool. The target-test set remains fixed across budgets. HM and

[1] In 23 of 560 Country-Agnostic I cells, the target-support set exceeds the entire source pool (a small source country paired with a large target country, e.g. India→Italy); there the source pool is exhausted and the context consists of target-support records only. This applies identically to both models, so the paired comparison is unaffected.

TabPFN use identical source users, target users, support records, test records, features, and preprocessing in each paired comparison.

2.5 Models

Random Forest Hybrid Model. The Hybrid Model uses source-population data and target-support data, and fits a new Random Forest classifier for every repeat, task, geographic setting, and support budget.

TabPFN. We use TabPFN v3 as the tabular in-context learning model in all experiments [3], with two ensemble members (`n_estimators = 2`), a fixed random seed, and the pretraining row limit disabled (`ignore_pretraining_limits = True`), running on a single NVIDIA RTX 4090 GPU. The full source-population and target-support set is passed as the in-context training set in a single forward pass, without subsampling, truncation, or chunking, and the fixed target-test data are predicted as query examples. Realized context sizes range from approximately 2.6K to approximately 217.8K records depending on the geographic setting; the largest contexts exceed TabPFN's pretraining scale, which we note as a caveat rather than a tuned choice. Features are median-imputed using only context rows, matching the Hybrid Model preprocessing exactly. TabPFN performs no task-specific gradient updates and does not fit a new classifier for each personalization condition [5].[2]

Target-only Random Forest. For the target-only analysis, the Random Forest is trained from scratch using only the target user's labeled support data. No source-population data are used.

Target-only TabPFN. The target-only TabPFN model uses the same target-user support data as context and predicts the same held-out target-user test data. It also uses no source-population data and performs no task-specific gradient updates.

2.6 Geographic Settings

We evaluate six geographic configurations, following and extending the DiversityOne evaluation design [7]. Table 1 summarizes how source and target users are defined in each setting.

3 Results

3.1 RQ1: TabPFN versus the Hybrid Model

We compare TabPFN with the Random Forest Hybrid Model (HM) across five target-support budgets and six geographic settings. All comparisons are paired, using identical source data, target-support data, fixed target-test data, labels, preprocessing, and support budgets. We report ΔAUROC as TabPFN − HM.

Table 2 shows a setting-dependent pattern. Country-Agnostic I shows the most consistent advantage for TabPFN, with significant improvements for both tasks across all support budgets. Country-specific three-class prediction also shows significant improvements across all budgets, whereas its two-class gains and both Country-Agnostic II comparisons are not statistically resolved across countries. Continent-specific and multi-country results are reported descriptively because these settings provide insufficient independent geographic units for the same statistical test.

3.2 RQ2: Target-Only Personalization

Table 3 shows that target-only TabPFN generally outperforms a Random Forest trained on the same user-specific support data across support budgets and geographic settings. In the country-specific evaluation, TabPFN outperforms RF in 71 of 80 country × task × budget comparisons. The pattern also persists under broader target-population pooling: TabPFN wins all 20 continent-specific continent × task × budget comparisons, all 10 natural multi-country task × budget comparisons, and 9 of 10 balanced multi-country comparisons. The only reversal occurs for balanced multi-country two-class prediction at 10% support.

4 Discussion

4.1 Main Findings

TabPFN is a strong, but not uniformly superior, training-free personalization baseline for DiversityOne mood inference. Its most consistent advantages appear in Country-Agnostic I, country-specific three-class prediction, and target-only personalization. In contrast, HM remains competitive or stronger in several pooled three-class settings, including natural multi-country pooling and higher-budget Country-Agnostic II. Because both models use identical source data, target-support data, fixed target-test data, features, and preprocessing, these differences primarily reflect the personalization mechanism: HM fits a task-specific Random Forest, whereas TabPFN conditions on the same examples in context [5].

Because the target-test set is fixed within each repeat, changes across 10%, 20%, 30%, 40%, and 50% support reflect the effect of adding more labeled target-user data. TabPFN is tied with HM at the smallest country-specific two-class budget but improves from 20% onward. In Country-Agnostic I, TabPFN outperforms HM at every budget for both tasks. In Country-Agnostic II, TabPFN improves two-class prediction across all budgets, while three-class prediction becomes slightly worse than HM at higher budgets. These patterns show that TabPFN's advantage depends on support budget, task formulation, and geographic setting, rather than representing a uniform improvement over HM.

As a temporal robustness check, we repeated the country-specific comparison using each target user's earlier 50% of records for personalization and later 50% for testing. At the 50% support budget, TabPFN remained ahead of HM for both two-class (0.628 vs. 0.581; $\Delta = +0.047$) and three-class prediction (0.657 vs. 0.635; $\Delta = +0.022$), with positive aggregate differences across all five budgets.

One possible explanation for the weaker TabPFN performance in some three-class settings is that three-class valence prediction is more sensitive to class imbalance and sparse neutral-class examples. In pooled geographic settings, the source context may also combine heterogeneous country-specific relationships between sensing features and mood labels. Under these conditions, fitting a task-specific Random Forest may remain competitive, especially when limited target-user support examples are insufficient to resolve both class imbalance and cross-country heterogeneity.

[2] We use two ensemble members rather than the default eight for tractability across the full sweep. On a three-country country-specific subset (30 cells, identical splits), raising the ensemble to eight changes macro AUROC by 0.010 on average (max 0.030): two-class results are essentially unchanged (−0.002 mean) and three-class results improve slightly (+0.010 mean). Our setting is therefore conservative with respect to the TabPFN–HM comparison.

Table 1: Geographic configurations used in our evaluation. Source and target users are disjoint under the user-level group 5-fold protocol unless the setting explicitly uses different source and target countries.

Setting	Source population	Target users	Notes
Country-specific	Users from one country	Held-out users from the same country	Per country
Continent-specific	Users pooled within Europe or Asia	Held-out users from the same pooled continent	No Americas
Country-Agnostic I	Users from one source country	Users from a different target country	Ordered pairs
Country-Agnostic II	Users pooled from seven countries	Users from the remaining held-out country	Leave-one-country-out
Multi-country natural	Users pooled from all eight countries, preserving original data volumes	Held-out users from the pooled multi-country dataset	Natural volume
Multi-country balanced	Users pooled from all eight countries after country-level downsampling	Held-out users from the balanced pooled dataset	Balanced volume

Table 2: RQ1 results across target-support budgets. Values are ΔAUROC = TabPFN − HM. Positive values indicate that TabPFN outperforms the Random Forest Hybrid Model. Bold values indicate statistically significant paired differences (two-sided Wilcoxon signed-rank test, $p < 0.05$). † denotes settings not tested due to insufficient independent geographic evaluation units.

	Two-class ΔAUROC					Three-class ΔAUROC				
Geographic setting	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
Country-specific	-0.001	+0.029	+0.036	+0.028	+0.030	**+0.023**	**+0.029**	**+0.030**	**+0.033**	**+0.034**
Continent-specific†	+0.025	+0.025	+0.034	+0.049	+0.042	-0.006	+0.001	-0.010	-0.008	-0.003
Country-Agnostic I	**+0.022**	**+0.022**	**+0.027**	**+0.029**	**+0.031**	**+0.026**	**+0.028**	**+0.028**	**+0.028**	**+0.027**
Country-Agnostic II	+0.024	+0.018	+0.022	+0.019	+0.011	+0.007	+0.004	+0.006	-0.006	-0.008
Multi-country natural†	+0.015	+0.022	+0.025	+0.026	+0.029	-0.010	-0.007	-0.014	-0.019	-0.021
Multi-country balanced†	+0.002	+0.006	+0.018	+0.011	+0.023	+0.023	+0.018	+0.017	+0.020	+0.025

Statistical tests use paired geographic evaluation units. At the 50% support budget, all bold comparisons remain significant after Holm–Bonferroni correction. Continent-specific and multi-country settings are reported descriptively because they contain fewer than three independent geographic units.

Table 3: RQ2 target-only personalization results across target-support budgets. Values are ΔAUROC = TabPFN − RF. Positive values indicate that TabPFN outperforms the target-only Random Forest trained on the same user-specific support data. Bold values indicate statistically significant paired differences (two-sided Wilcoxon signed-rank test, $p < 0.05$). † denotes settings not tested due to insufficient independent geographic evaluation units.

	Two-class ΔAUROC					Three-class ΔAUROC				
Geographic setting	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
Country-specific	+0.008	**+0.033**	**+0.034**	**+0.029**	**+0.036**	**+0.017**	**+0.025**	**+0.029**	**+0.028**	**+0.023**
Continent-specific†	+0.027	+0.030	+0.038	+0.043	+0.040	+0.031	+0.033	+0.030	+0.031	+0.030
Multi-country natural†	+0.009	+0.023	+0.030	+0.033	+0.037	+0.020	+0.024	+0.023	+0.023	+0.021
Multi-country balanced†	-0.013	+0.005	+0.017	+0.024	+0.026	+0.007	+0.017	+0.024	+0.023	+0.025

Statistical tests use paired geographic evaluation units. At the 50% support budget, both country-specific comparisons remain significant after Holm–Bonferroni correction. Continent-specific and multi-country settings are reported descriptively because they contain fewer than three independent geographic units.

4.2 Relation to DiversityOne Previous Studies

The original generalization and personalization study showed that partial personalization improves mobile mood inference and that geographic structure affects generalization [7]. Our contribution is not to re-establish personalization, but to test whether TabPFN can use the same personalization data through in-context learning. We preserve the original geographic framing and further include a paired HM–TabPFN comparison for Country-Agnostic II under the leave-one-country-out fixed-test protocol.

4.3 Target-Only Personalization

The target-only experiment represents a source-free calibration scenario: a new user provides a labeled support set, and held-out records from that user are predicted using only those user-specific examples. TabPFN outperforms a target-only Random Forest without source-population data or task-specific gradient updates. This suggests that TabPFN can serve as a source-free personalization baseline.

Target-only TabPFN also performs strongly across support budgets and population-pooling settings. Although its observed AUROC is sometimes higher than in the source-augmented experiments, these settings are not a controlled paired comparison of source versus source-free context. The results therefore motivate future study of when source-population context helps or dilutes target-specific personalization.

4.4 Limitations and Future Work

This study is limited to one dataset, DiversityOne [1]. Our reconstructed HM reproduces the personalization effect but not every

absolute AUROC value, partly because our reconstruction applies different filtering and valid-window selection. In a target-only sensitivity analysis, removing the filtering reduced absolute AUROC for both models but changed the TabPFN–RF gap by at most 0.014, providing little evidence that filtering differentially favors TabPFN. At low support budgets, some users lack minority-class examples in their support sets, particularly for the imbalanced two-class formulation; we therefore interpret the 10% condition cautiously. TabPFN required 7.4× the total wall-clock time of HM in our country-specific benchmark; thus, "training-free" refers to the absence of task-specific parameter optimization rather than faster computation.

Future work should evaluate TabPFN on additional mobile and wearable sensing datasets, compare additional tabular foundation models, and study when source-population context improves or dilutes target-specific personalization.

Ethical Statement. This study uses the DiversityOne dataset under its approved research-use conditions. As mood inference may produce inaccurate or culturally biased predictions, these methods should not be used for high-stakes decisions without further validation and human oversight.

5 Conclusion

We evaluated TabPFN as a training-free personalization baseline for mobile mood inference on DiversityOne. Under a paired fixed-test protocol, its most consistent advantages over the Random Forest Hybrid Model appear in Country-Agnostic I and country-specific three-class prediction, while HM remains competitive or stronger in several pooled three-class settings. In target-only personalization, TabPFN also generally outperforms a Random Forest trained on the same user-specific support data across support budgets and geographic pooling strategies. A chronological robustness analysis further shows that the advantage persists when earlier target-user records are used to predict later records. Overall, these results position tabular in-context learning as a strong, but setting-dependent, training-free personalization approach for mobile affective and mental sensing.

References

[1] Matteo Busso, Andrea Bontempelli, Leonardo Javier Malcotti, Lakmal Meegahapola, Peter Kun, Shyam Diwakar, Chaitanya Nutakki, Marcelo Dario Rodas Britez, Hao Xu, Donglei Song, Salvador Ruiz Correa, Andrea-Rebeca Mendoza-Lara, George Gaskell, Sally Stares, Miriam Bidoglia, Amarsanaa Ganbold, Altangerel Chagnaa, Luca Cernuzzi, Alethia Hume, Ronald Chenu-Abente, Roy Alia Asiku, Ivan Kayongo, Daniel Gatica-Perez, Amalia de Götzen, Ivano Bison, and Fausto Giunchiglia. 2025. DiversityOne: A Multi-Country Smartphone Sensor Dataset for Everyday Life Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 1 (March 2025), 49 pages. doi:10.1145/3712289

[2] Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai "Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 56 (May 2024), 25 pages. doi:10.1145/3659604

[3] Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. 2026. TabPFN-3: Technical Report. arXiv:2605.13986 https://arxiv.org/abs/2605.13986

[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] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2023. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations*. https://openreview.net/forum?id=cp5PvcI6w8_

[6] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. In *Proceedings of the Fifth Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research, Vol. 248)*. PMLR, 522–539. https://proceedings.mlr.press/v248/kim24b.html

[7] 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

[8] Alexandre Nanchen, Lakmal Meegahapola, William Droz, and Daniel Gatica-Perez. 2023. Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 217–228. doi:10.1145/3600211.3604688

[9] 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

[10] 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

[11] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1, Article 31 (March 2024), 32 pages. doi:10.1145/3643540

[12] 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

[13] Tianyi Zhang, Songyan Teng, Hong Jia, and Simon D'Alfonso. 2024. Leveraging LLMs to Predict Affective States via Smartphone Sensor Features. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '24)*. Association for Computing Machinery, 709–716. doi:10.1145/3675094.3678420