DIVERSITYONE OPEN CHALLENGE '26

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

Tabular in-context learning vs. the Random Forest Hybrid Model on DiversityOne

Panyu Zhang* · panyu@kaist.ac.kr
with Shohruh Shokulov*, Humoyunbek Abdukarimov, Jumabek Alikhanov, Surjya Ghosh, Uichin Lee†
KAIST · HumbleBeeAI · Gachon University · BITS Pilani, K K Birla Goa Campus (* equal contribution · † corresponding author)

Personalization Helps, but Requires Re-fitting

Population-level mood models generalize poorly to unseen users and countries. A few labeled examples from the target user help—but conventional personalization requires fitting a new classifier for each setting.

- Behavioral and sensing distributions vary across users and countries
- The Random Forest Hybrid Model reduces this gap by combining source-population data with labeled target-user examples
- But each new user, source population, or personalization budget requires fitting a new Random Forest

Observed
Population-level models degrade on unseen users and countries

Costly
A new task-specific classifier is fitted for each user, source population, and personalization budget

Question
Can we use the same target-user labels without task-specific model fitting?

Meegahapola, L., et al. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. Proc. ACM IMWUT, 2023.
Xu, X., et al. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. Proc. ACM IMWUT, 2023.

CORE IDEA

Training-Free Personalization with TabPFN

We keep the personalization data and test set fixed, and change only how the labeled examples are used.

Random Forest Hybrid Model

Meegahapola et al., 2023 · the DiversityOne personalization baseline

Source-population rows + labeled target-user support rows

↓

Fit a new Random Forest

↓

Predictions for the held-out target-user test rows

Refit as the user, source population, or support budget changes

TabPFN

Pretrained for tabular prediction; conditions on labeled examples at inference time (Hollmann et al., 2023)

Source-population rows + labeled target-user support rows

↓

Use as in-context examples

↓

Predictions for the held-out target-user test rows

Same pretrained model · no task-specific parameter updates

Hollmann, N., et al. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. ICLR, 2023.
Grinsztajn, L., et al. TabPFN-3: Technical Report. arXiv:2605.13986, 2026.

RESEARCH QUESTIONS

Two Questions About Training-Free Personalization

We first test TabPFN with source-population data, then ask whether it can personalize from target-user labels alone.

RQ1 — How does TabPFN compare with the Random Forest Hybrid Model?

Same source data, target-user support, and fixed test records across 5 budgets × 6 geographic settings × 2 valence tasks.

RQ2 — Can TabPFN personalize using only the target user's labeled data?

No source-population data. TabPFN and the Random Forest use the same target-user support and test records.

Paired evaluation

Identical data and preprocessing within each RQ comparison; only the personalization mechanism differs.

Meegahapola, L., et al. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. Proc. ACM IMWUT, 2023.

DiversityOne: Personalization Across 8 Countries

We reconstruct DiversityOne from its public raw sensing release and predict time-diary mood valence from features aggregated within ±5 minutes of each report.

- **Two-class:** valence 1–3 positive vs. 4–5 negative
- **Three-class:** 1–2 positive, 3 neutral, 4–5 negative
- **Primary metric**: macro AUROC (one-vs-rest for three classes)

Public-release reconstruction: not an exact reproduction of the original feature table (sanity check in backup).

8

countries: China, Denmark, India, Italy, Mexico, Mongolia, Paraguay, UK

526

users in the final evaluated pool

221k

reconstructed valid sensing windows

~102

features: location, connectivity, screen, activity, apps ...

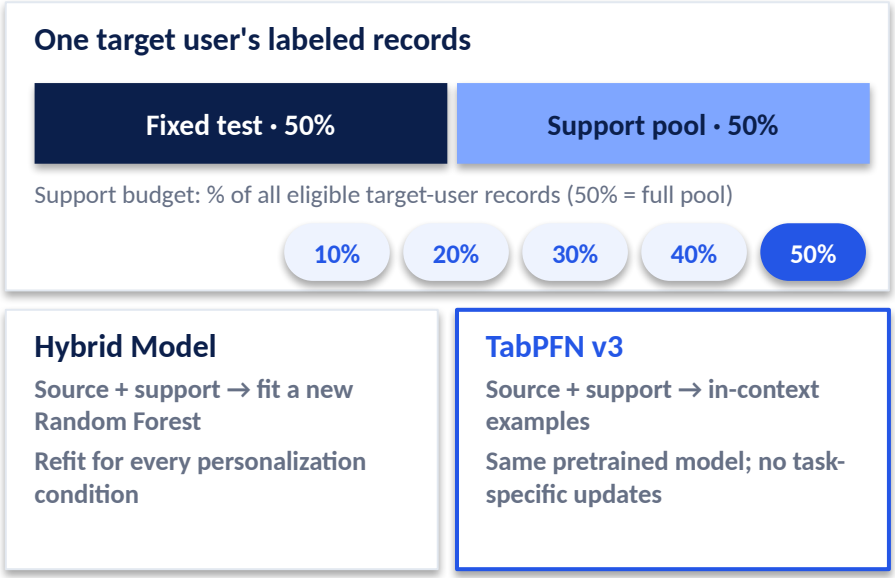
Busso, M., et al. DiversityOne: A Multi-Country Smartphone Sensor Dataset for Everyday Life Behavior Modeling. Proc. ACM IMWUT, 2025.
Meegahapola, L., et al. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. Proc. ACM IMWUT, 2023.

PROTOCOL

A Paired, Fixed-Test Target-Support Protocol

Each target user keeps a fixed 50% test set. At every label budget, both models receive exactly the same records.

- **Fixed test**: stratified 50% of each target user's records; the remaining 50% forms the support pool
- **Support budgets**: 10%, 20%, 30%, 40%, 50% of all eligible records; 50% uses the full support pool
- **Fixed context size**: each added support row replaces one randomly sampled source row



Same source rows + same support rows + same fixed test rows

→ only the personalization mechanism changes

Meegahapola, L., et al. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. Proc. ACM IMWUT, 2023.

Grinsztajn, L., et al. TabPFN-3: Technical Report. arXiv:2605.13986, 2026.

PROTOCOL

Six Geographic Settings from DiversityOne

Source and target users are always different people. Within-country settings test new users; the agnostic and pooled settings test new countries.

Setting	Source population	Target users	Notes
Country-specific	One country	Held-out users, same country	Per country
Continent-specific	Pooled Europe or Asia	Held-out users, same continent	No Americas
Country-Agnostic I	One source country	Users of a different country	Ordered pairs
Country-Agnostic II	Seven countries pooled	The remaining held-out country	Leave-one-country-out
Multi-country (natural)	All eight, original volumes	Held-out users of the pool	Natural volume
Multi-country (balanced)	All eight, downsampled per country	Held-out users of the pool	Balanced volume

Within-country settings test personalization to new users; the agnostic and pooled settings test transfer to new countries and the effect of mixed-country source data.

Meegahapola, L., et al. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. Proc. ACM IMWUT, 2023.

TabPFN Is Competitive, with Setting-Dependent Gains

The strongest TabPFN gains appear in Country-Agnostic I and country-specific three-class prediction; HM wins only in natural multi-country three-class from 20% onward.

	Two-class ΔAUROC					Three-class ΔAUROC				
Budget →	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**

ΔAUROC = TabPFN − HM. Outlined + bold: $p < .05$, paired Wilcoxon over 10 repeated holdouts.

Meegahapola, L., et al. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. Proc. ACM IMWUT, 2023.

TabPFN Personalizes Effectively From Target-User Labels Alone

With no source-population data, TabPFN significantly outperforms a Random Forest trained on the same target-user records in 33 of 40 conditions.

Budget →	Two-class ΔAUROC 10%	20%	30%	40%	50%	Three-class ΔAUROC 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**

ΔAUROC = TabPFN − target-only RF. Outlined + bold: $p < .05$ (paired Wilcoxon over 10 holdouts). Country-Agnostic settings have no target-only analogue.

RESULTS · ROBUSTNESS CHECK

Does the Advantage Hold Under a Chronological Split?

We replace the random within-user split with a chronological one: each user's earlier 50% forms the personalization pool, and the later 50% is held out for testing.

Two-class · 50% support

TabPFN	Hybrid Model	Difference
0.628	0.581	✦0.047

Macro AUROC · earlier half → personalization, later half → test

Three-class · 50% support

TabPFN	Hybrid Model	Difference
0.657	0.635	✦0.022

Macro AUROC · earlier half → personalization, later half → test

The advantage persists when personalization precedes testing in time

TabPFN remains ahead at all five support budgets (10–50%); at 50% support, ΔAUROC = +0.047 / +0.022 for two-/three-class prediction.

DISCUSSION

A Strong but Setting-Dependent Baseline

Because the paired comparison fixes the data and preprocessing, performance differences primarily reflect how the same labeled examples are used: HM fits a new forest, while TabPFN conditions on them in context.

- TabPFN is ahead in 48/60 RQ1 conditions, significantly so in 36; HM is significantly ahead in only 4.
- The advantage persists under a chronological split: TabPFN remains ahead at all five budgets when personalization uses earlier records and testing uses later records.
- HM retains a significant advantage in natural multi-country three-class prediction from 20% support onward.
- Target-only TabPFN can sometimes match or exceed source-augmented performance, suggesting that more source context is not always beneficial.

Training-free personalization is viable

A pretrained tabular model can use limited target-user labels competitively without fitting a task-specific classifier.

But gains depend on the setting

Natural multi-country pooling in the three-class task favors the retrained forest.

Open question

When does source-population context help, and when does it dilute the target user's own signal?

Limitations and Next Steps

What this study does not show, and what we plan to do next.

Next steps

- Evaluate on additional mobile-sensing datasets
- Benchmark additional tabular foundation models
- Study when source-population context helps—and when it dilutes target-specific personalization

Limited external validity

Evaluated only on DiversityOne. Additional datasets are needed to establish generalizability.

Training-free ≠ computationally cheaper

TabPFN required 7.4× HM's wall-clock time; contexts up to ≈218K rows also exceed its pretraining scale.

Caution at low support

At 10% support, some users have few or no minority-class examples, particularly in the two-class task.

Busso, M., et al. DiversityOne: A Multi-Country Smartphone Sensor Dataset for Everyday Life Behavior Modeling. Proc. ACM IMWUT, 2025.
Grinsztajn, L., et al. TabPFN-3: Technical Report. arXiv:2605.13986, 2026.

DIVERSITYONE OPEN CHALLENGE ’26

Thank You — Questions?

Panyu Zhang · KAIST · panyu@kaist.ac.kr

Paper

Code & results

Take-home message

A pretrained tabular model can personalize mood inference from a few target-user labels, with no retraining, and matches or beats the retrained Random Forest Hybrid Model in most settings.

Gains depend on the setting: strongest in Country-Agnostic I, country-specific three-class and target-only personalization; the Hybrid Model still leads in natural pooled three-class prediction.