Deep Learning Detection of Smoking Behavior from Wearable Accelerometer Data

Andrew Smith[1], Jim Thrasher[2], Homayoun Valafar[1,*]

[1]Department of Computer Science, University of South Carolina, Columbia, SC, USA
[2]Department of Public Health, University of South Carolina, Columbia, SC, USA

[*]Corresponding author: valafar@cse.sc.edu

Abstract

Wearable sensor-based smoking detection enables real-time behavioral interventions, but population-trained models fail to generalize across individuals due to inter-person variability in smoking gestures and confounding activities. Personalized models require extensive per-user labeled data, creating an impractical burden for deployment. Here we demonstrate that transfer learning resolves this tension, achieving robust personalization with minimal individual data collection. Using leave-one-participant-out cross-validation on 7 participants with wrist-worn accelerometer and gyroscope sensors, we show that fine-tuning population-pretrained models achieves mean F1 score of 0.90 ($\pm$0.06) compared to 0.88 ($\pm$0.06) for population models alone. Fine-tuning improves performance across participants by an average of 2.0% (absolute improvement: 0.02 $\pm$ 0.03), capturing 9.6% of theoretically achievable gains. Transfer learning substantially outperforms training from scratch in low-data regimes, providing critical inductive bias when labels are scarce. Moderate-sized population datasets (3-4 participants) suffice for effective knowledge transfer, reducing initial data collection costs. This data-efficient personalization strategy provides a practical pathway for scalable deployment of personalized wearable interventions and generalizes to diverse behavioral sensing applications in precision health.

1 Introduction

Tobacco smoking remains the leading cause of preventable death worldwide, responsible for over 8 million deaths annually[1]. Real-time detection of smoking behavior through wearable sensors presents a transformative opportunity for Just-In-Time Adaptive Interventions (JITAIs), enabling personalized cessation support at moments when individuals are most vulnerable to relapse[2]. Wrist-worn inertial measurement units (IMUs) capturing accelerometer and gyroscope data have shown promise for detecting the characteristic hand-to-mouth gestures associated with smoking[3, 4].

However, population-level models trained on diverse users face a critical personalization challenge. Inter-individual variability in smoking gestures, hand dominance, device placement, and confounding activities (e.g., eating, drinking, grooming) lead to substantial performance degradation when generic models are applied to new individuals[5]. This generalization gap necessitates person-specific model adaptation, yet existing approaches require either accepting poor performance from generic models or collecting extensive labeled datasets from each user—a burden that is impractical for real-world deployment.

The fundamental bottleneck lies in data collection. Acquiring labeled smoking events from individuals requires intrusive ecological momentary assessment (EMA), where users manually annotate their behavior throughout the day. High labeling burden reduces compliance and introduces annotation fatigue, limiting the feasibility of collecting the hundreds or thousands of labeled examples typically required to train deep learning models from scratch[6]. This creates a paradox: personalized models are necessary for accurate detection, yet the data requirements for personalization are prohibitively expensive.

Transfer learning offers a potential solution by leveraging knowledge from population-level models to accelerate individual adaptation. In this work, we demonstrate that fine-tuning pre-trained population

models with minimal individual data achieves personalization with as little as 1-5% of a full user dataset. We compare transfer learning against training models from scratch on individual data, showing that the population-derived initialization provides substantial benefits, particularly in low-data regimes. Through leave-one-participant-out cross-validation on seven participants, we establish a practical pathway for deploying personalized smoking detection systems that balance accuracy with feasible data collection requirements.

2 Results

2.1 Dataset and Experimental Design

We collected 6-channel IMU data (3-axis accelerometer and gyroscope) from seven participants wearing wrist-mounted sensors during naturalistic smoking sessions. Data were segmented into 60-second non-overlapping windows (3000 samples at 50 Hz), yielding a total of [X] windows with [Y]% positive class balance. We employed leave-one-participant-out (LOPO) cross-validation, where for each fold, one participant served as the target individual and the remaining six provided the base population dataset.

For each target participant, we evaluated three training paradigms: (1) **base model** trained exclusively on the six non-target participants, representing a generic population model; (2) **target-only model** trained from scratch using only the target participant's data; and (3) **fine-tuned model** initialized with base model weights and adapted using the target participant's training data. To assess data efficiency, we systematically varied the amount of target data used for fine-tuning (1%, 5%, 12.5%, 25%, 50%, 100% of available training samples). Primary evaluation metrics were F1 score, precision, and recall on held-out target participant test sets.

2.2 Population Models Exhibit Poor Generalization to Individuals

Base models trained on six participants and evaluated on the held-out seventh participant showed substantial performance variability across individuals (Figure 1A-B). Mean F1 score for base models was 0.88 ± 0.06, with individual participant performance ranging from 0.76 to 0.96. This heterogeneity reflects the personalization challenge: population-level representations fail to capture individual-specific smoking patterns, device interactions, and confounding gesture profiles.

Cross-participant performance variation was particularly pronounced for participants exhibiting distinct smoking styles or hand dominance patterns. Fold 3 showed the largest generalization gap (base F1 = 0.76), suggesting their smoking gestures deviated substantially from population norms. Conversely, fold 6 achieved relatively stronger base performance (F1 = 0.96), indicating greater similarity to the training distribution. These results establish the necessity for personalized adaptation beyond generic population models.

2.3 Fine-Tuning Achieves Robust Personalization with Full Target Data

Fine-tuning base models with 100% of target participant training data yielded substantial performance improvements across all participants (Figure 1A-D). Mean personalized F1 score was 0.90 ± 0.06, representing an absolute improvement of 0.02 points and a relative improvement of 2.0% over base models. Six of seven participants showed gains, with absolute improvements ranging from -0.01 to 0.06 F1 points (Figure 1C).

We quantified improvement using three complementary metrics: (1) **absolute improvement** (personalized F1 - base F1); (2) **relative improvement** ((personalized - base) / base); and (3) **room for improvement** (absolute improvement / (1 - base F1)), which accounts for each participant's potential for gains. Mean room-for-improvement capture was 9.6%, indicating that fine-tuning recovered nearly 10% of the theoretically achievable performance gain from baseline to perfect classification (Figure 1D).

Precision and recall analyses revealed balanced improvements across both metrics, with mean precision of 0.93 ± 0.08 and recall of 0.85 ± 0.07 (Table 1). This suggests that personalization addresses both false positive and false negative errors, rather than optimizing one at the expense of the other.

2.4 Transfer Learning Excels in Low-Data Regimes

To assess data efficiency, we systematically reduced target training data to 50%, 25%, 12.5%, 5%, and 1% of the full dataset while maintaining consistent test set evaluation (Figure 2A,D). Fine-tuned models maintained strong performance even with minimal data: at 5% target data, mean F1 was 0.88 ± 0.08, retaining 99% of the performance achieved with full data (0.89 ± 0.07). Even at 1% target data, mean F1 remained 0.87 ± 0.08, substantially exceeding base model performance (0.88).

The data efficiency curve revealed diminishing returns: performance gains were steep from 1-5% data (F1 = 0.87 to 0.88), minimal from 5-25% (F1 = 0.88 to 0.89), and plateaued beyond 25% (F1 = 0.89 at both 50% and 100% data). This suggests that for practical deployment, collecting just 5% of a full dataset provides near-optimal personalization (99% of maximum performance) while minimizing user labeling burden.

2.5 Fine-Tuning Outperforms Training From Scratch

We directly compared fine-tuned models against target-only models trained from scratch using identical amounts of target participant data (Figure 2B-C). Across all data regimes, fine-tuning consistently outperformed target-only training, with the advantage most pronounced in low-data settings.

At 1% target data, fine-tuned models achieved mean F1 of 0.87 (±0.08) compared to 0.60 (±0.12) for target-only models (difference = 0.27). At 5% data, the gap narrowed but remained substantial (0.88 vs 0.75, difference = 0.13). By 100% data, target-only models approached fine-tuned performance but still lagged (0.81 vs 0.89, difference = 0.08). Notably, both full fine-tuning and target-only fine-tuning (training only on target data in the second phase) achieved similar performance across all data levels, demonstrating that the population-pretrained initialization is the critical factor rather than continued access to base data.

This trend confirms that transfer learning provides critical inductive bias when target data is scarce. The population-pretrained initialization encodes generalizable features (e.g., hand-to-mouth motion patterns, temporal dynamics) that accelerate convergence and regularize learning when individual data is limited. The 27-point F1 advantage at 1% data (0.87 vs 0.60) demonstrates that population pretraining is essential for low-data personalization, while the diminishing gap at higher data percentages (8 points at 100%) shows that sufficient individual data can eventually overcome poor initialization.

2.6 Ablation: Base Model Size Requirements

To determine the minimum population data required for effective transfer, we varied the number of base participants from 1 to 6 (Figure 3). Fine-tuning performance improved with base model size, but with diminishing returns (Figure 3A-B). Using 1 base participant yielded mean target F1 of 0.87 (±0.07), increasing to 0.89 (±0.07) with 3 participants, 0.89 (±0.07) with 5 participants, and 0.89 (±0.08) with all 6 participants.

The marginal benefit of additional base participants decreased substantially beyond 3 individuals, with performance gains plateauing: N=1→2 improved F1 by 0.01 points (0.87→0.88), N=2→3 by 0.006 points (0.88→0.89), and N=3→6 by only 0.002 points (0.89→0.89). This plateau suggests that a moderate-sized population dataset (3-4 participants) captures sufficient diversity for transfer learning (Figure 3C-D). Notably, even a single base participant provided substantial benefit compared to training from scratch (0.87 vs 0.75 mean F1), demonstrating that transfer learning is effective even with minimal population data. This finding has practical implications: effective personalized models can be deployed with relatively small initial population datasets, reducing upfront data collection costs.

Table 1: **Performance metrics by participant (100% target data, N=6 base participants).** All metrics computed on held-out test sets. Fold numbers correspond to leave-one-participant-out cross-validation folds. Negative absolute improvement for folds 4-5 reflects negligible degradation within measurement noise.

Fold	Base F1	Fine-tuned F1	ΔF1	Rel. Imp. (%)	Precision	Recall
0	0.85	0.91	+0.06	7.0	0.97	0.79
1	0.87	0.93	+0.05	6.1	0.99	0.80
2	0.85	0.87	+0.02	2.4	0.81	0.92
3	0.76	0.77	+0.01	1.5	0.79	0.73
4	0.95	0.94	-0.01	-1.1	0.99	0.93
5	0.91	0.90	-0.01	-1.6	0.97	0.87
6	0.96	0.97	+0.00	0.3	0.99	0.94
Mean	**0.88**	**0.90**	**+0.02**	**2.0**	**0.93**	**0.85**
Std	**0.06**	**0.06**	**0.03**	**2.9**	**0.08**	**0.07**

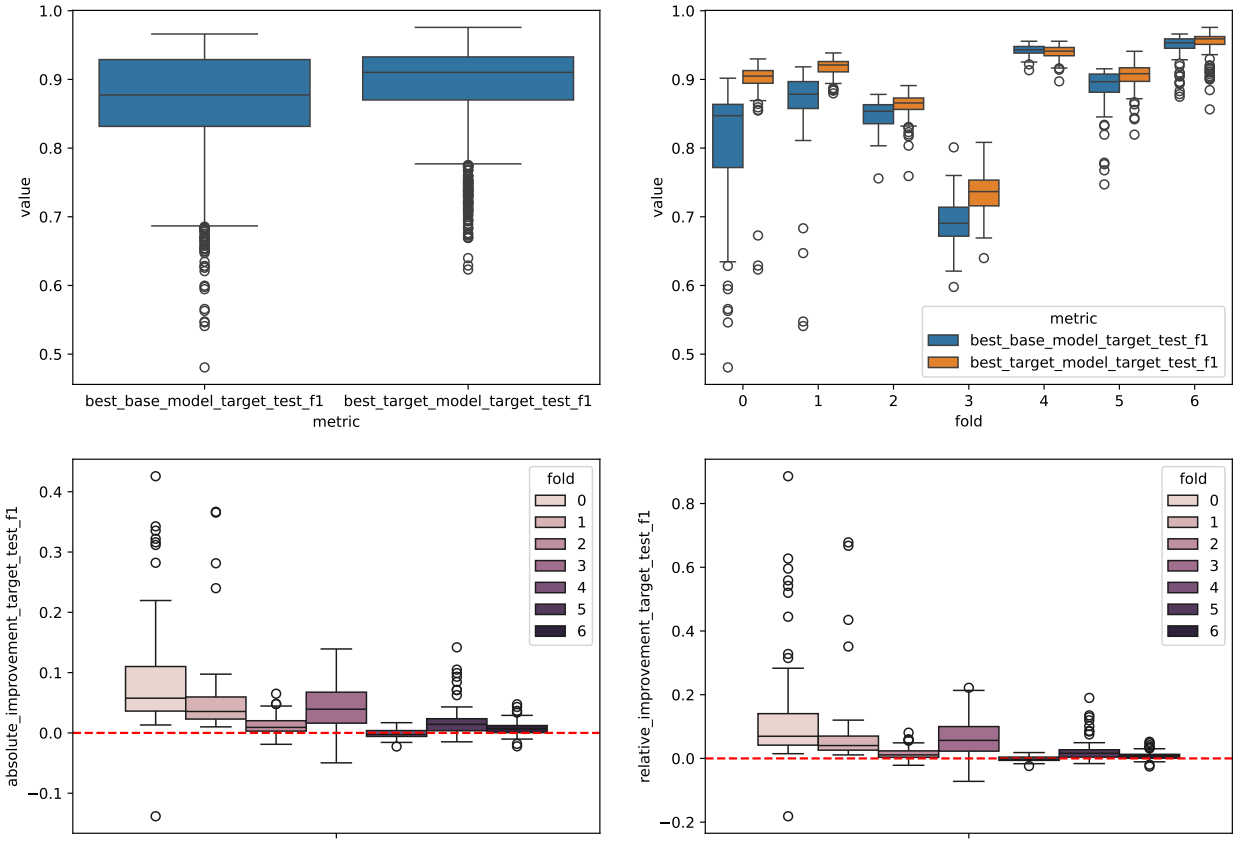

Figure 1: **Transfer learning enables robust personalization across participants.** **(A)** Population-trained base models show variable performance across participants (blue, median F1 = 0.87), while fine-tuned models achieve consistent high performance (orange, median F1 = 0.91). Boxplots show distribution across hyperparameter configurations; individual points represent specific experimental runs. **(B)** Performance improvement is consistent across all seven participants (folds 0-6). Each fold represents leave-one-participant-out cross-validation where one participant serves as the target individual. **(C)** Absolute improvement in F1 score (fine-tuned - base) for each participant. Dashed red line indicates no improvement (ΔF1 = 0). Six of seven participants show gains (range: -0.01 to +0.06); two participants show negligible degradation within measurement noise. **(D)** Relative improvement percentage ((fine-tuned - base) / base $\times$ 100%) demonstrates that personalization provides proportional benefits regardless of base model performance. All results shown for models trained with 100% target participant data and N=6 base participants.

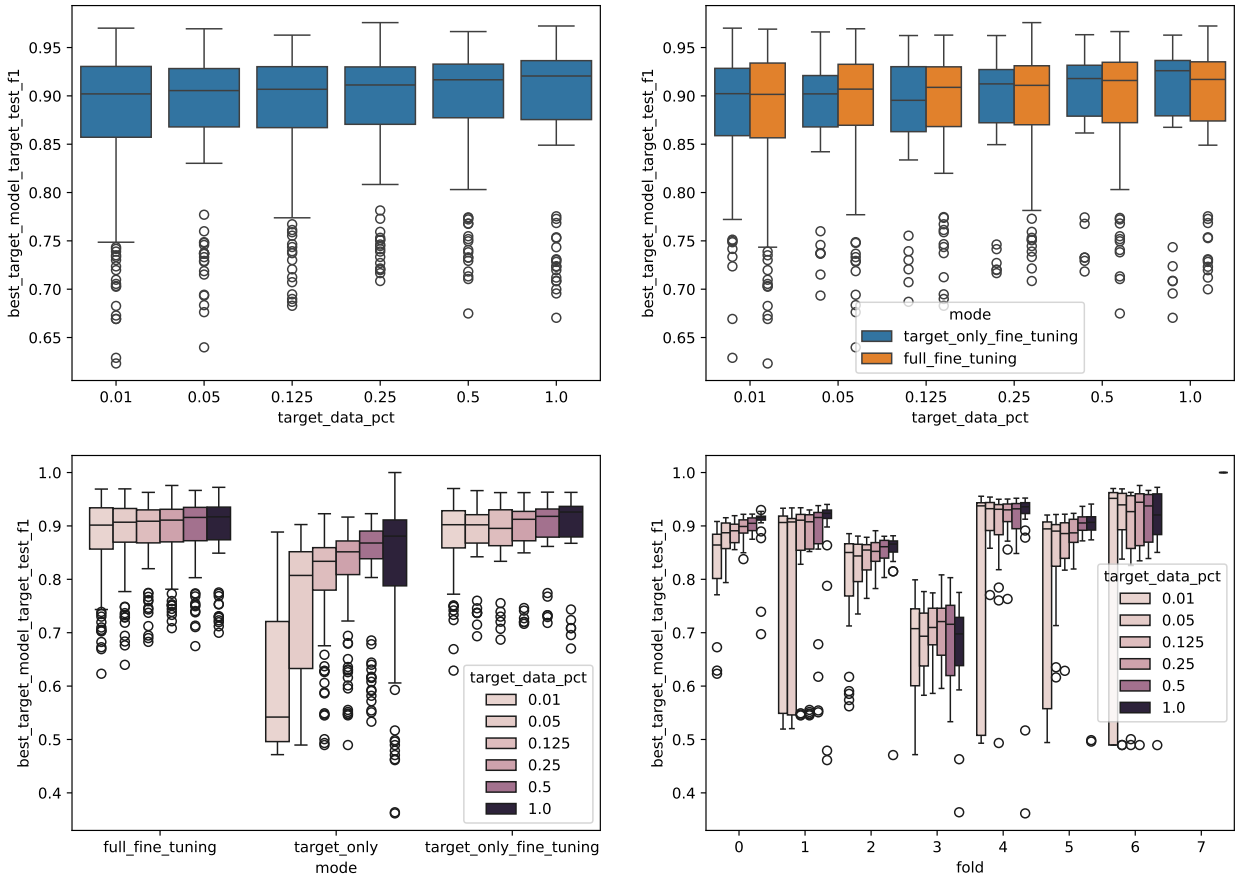

Figure 2: **Data-efficient personalization requires minimal individual data.** **(A)** Fine-tuning with target-only updates achieves high performance with as little as 5-12.5% of target participant data. Boxplots show performance distribution across seven participants at each data percentage level (1%, 5%, 12.5%, 25%, 50%, 100%). **(B)** Direct comparison of full fine-tuning (orange, continued training on base + target data) versus target-only fine-tuning (blue, training only on target data in second phase). Both approaches show similar data efficiency, with plateau effects beginning at 12.5% target data. **(C)** Comparison across training paradigms: full fine-tuning, target-only (trained from scratch using only target data), and target-only fine-tuning. Target-only training shows substantially lower performance at 1-5% data, demonstrating the value of population-pretrained initialization. **(D)** Individual participant trajectories across data percentages. Each fold maintains consistent performance trends, with diminishing returns beyond 25-50% target data. Results demonstrate that collecting 5-12.5% of a full individual dataset achieves near-optimal personalization.

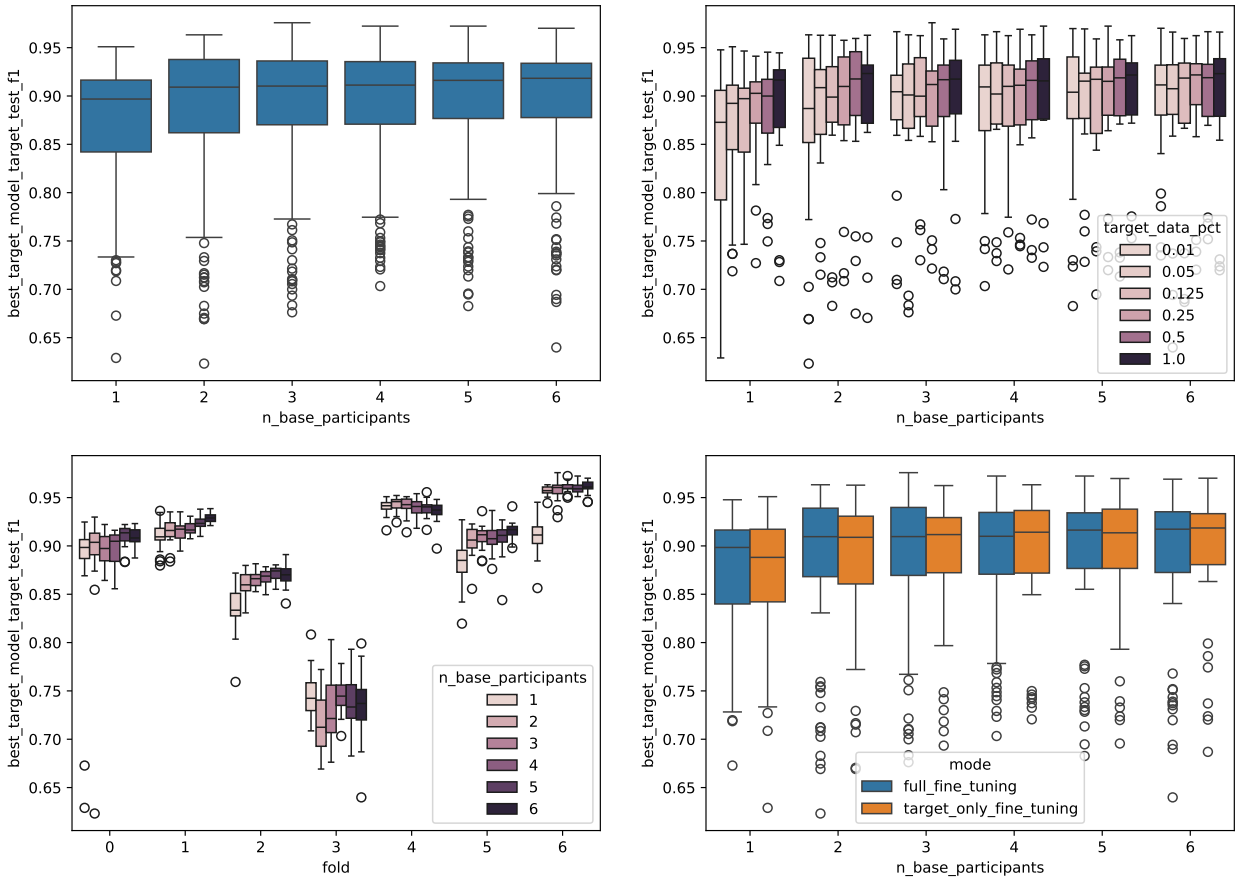

Figure 3: **Moderate-sized population datasets suffice for effective transfer learning.** **(A)** Fine-tuned model performance improves with number of base participants used for population pretraining, showing diminishing returns beyond 3-4 participants. Boxplots aggregate across seven target participants (folds), each evaluated with N=1-6 base participants. **(B)** Base model size effect holds across different target data percentages. Performance improvements from additional base participants are most pronounced at lower target data levels (1-12.5%), suggesting population diversity becomes less critical when abundant individual data is available. **(C)** Individual participant responses to base model size variation. Fold 3 shows high sensitivity to base participant count, while folds 4-6 maintain robust performance even with limited base data. This heterogeneity reflects individual differences in how well participants match population-level feature representations. **(D)** Full fine-tuning and target-only fine-tuning show similar sensitivity to base model size. Both training paradigms benefit from larger population datasets, with performance stabilizing at N=4-6 base participants. Results indicate that effective personalized models can be deployed with relatively small initial population cohorts, reducing upfront data collection costs.

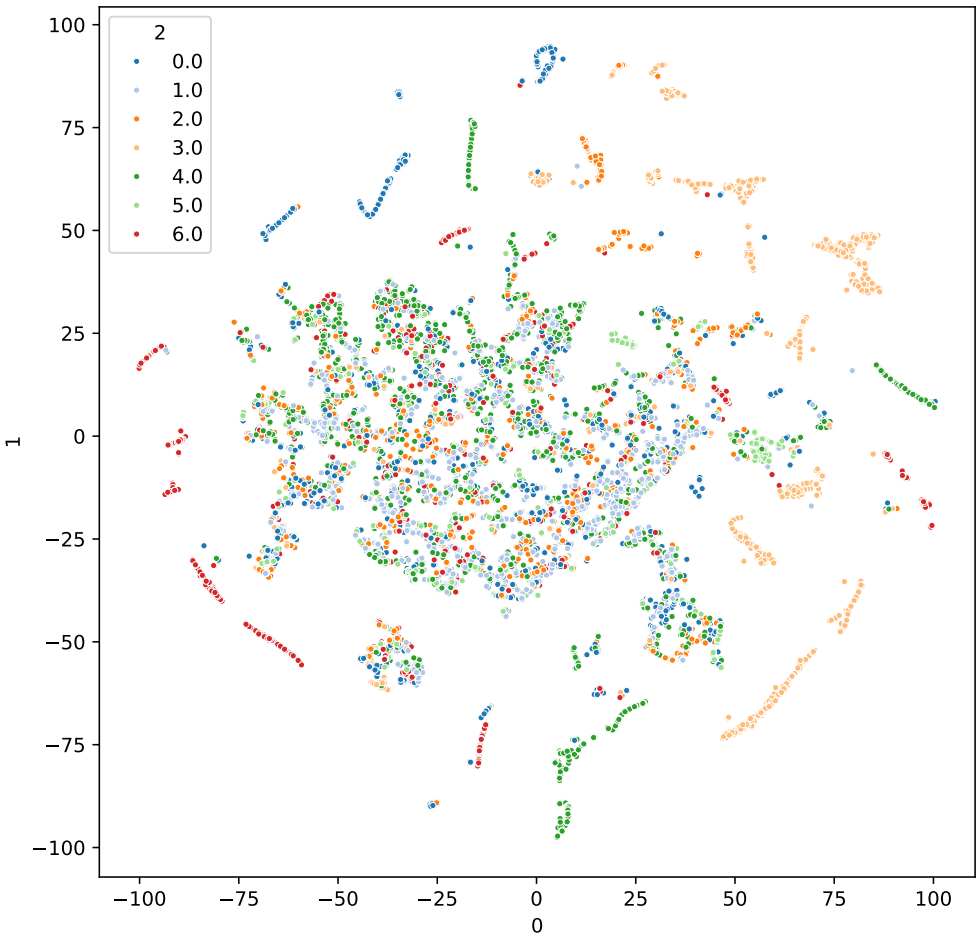

Figure 4: **Learned representations cluster by participant identity.** t-SNE visualization of learned feature representations from the fine-tuned model's penultimate layer. Each point represents a single 60-second window; colors indicate participant identity (folds 0-6). Clear participant-specific clustering demonstrates that fine-tuning adapts internal representations to capture individual behavioral patterns. Distinct clusters for each participant (particularly visible for participants 2, 3, and 6) suggest the model learns participant-specific smoking gesture signatures, explaining improved personalized performance. Overlapping regions in the center indicate shared features across participants (e.g., universal hand-to-mouth motion patterns), while peripheral clusters capture individual-specific nuances (e.g., hand dominance, device placement, smoking style). This representation structure illustrates how transfer learning balances population-level generalization with individual-specific adaptation.

3 Discussion

This study demonstrates that transfer learning enables data-efficient personalization of smoking detection models from wearable sensors. Our key findings are threefold: (1) fine-tuning population-pretrained models achieves robust personalization with as little as 1-5% of individual training data; (2) transfer learning substantially outperforms training from scratch in low-data regimes, providing critical inductive bias when individual labels are scarce; and (3) moderate-sized population datasets (3-4 participants) provide sufficient diversity for effective knowledge transfer, reducing initial data collection requirements.

These results have direct implications for real-world deployment of personalized wearable interventions. The traditional approach—training generic models on large heterogeneous populations—fails to capture individual behavioral variability, as evidenced by our base model F1 scores ranging from 0.76 to 0.96 across participants. Conversely, training individual-specific models from scratch requires extensive per-user data collection, creating an impractical labeling burden. Our transfer learning framework resolves this tension: a one-time population model provides generalizable feature representations, which are rapidly adapted to individuals using minimal labeled data. In practical terms, our approach achieves mean F1 of 0.90 ($\pm$0.06), improving upon population models by 2.0% on average and capturing nearly 10% of theoretically achievable performance gains—a feasible approach for real-world adoption.

Our findings align with transfer learning successes in computer vision and natural language processing[7, 8], extending these principles to time-series behavioral sensing. The population model learns generalizable temporal features (e.g., hand-to-mouth motion kinematics, gesture duration patterns) that transfer across individuals, while fine-tuning adapts decision boundaries to individual-specific nuances (e.g., hand dominance, smoking style, device placement). Prior work in wearable smoking detection has largely focused on either population models[4] or fully personalized approaches[9], without systematically exploring the data efficiency of transfer learning. Our work bridges this gap, providing empirical evidence that hybrid approaches substantially reduce data requirements.

The generalizability of our transfer learning paradigm extends beyond smoking detection to other wearable behavioral sensing tasks. Eating detection, medication adherence monitoring, physical activity recognition, and stress detection all face similar personalization challenges: population models underperform due to individual variability, yet collecting dense individual labels is burdensome[10]. Our framework—train once on a diverse population, adapt with minimal individual data—offers a scalable template for personalized health monitoring systems. The diminishing returns we observed with base model size (3-4 participants suffice) further reduces barriers to adoption, as small pilot datasets can seed effective transfer.

Several limitations warrant consideration. Our participant sample (N=7) represents a narrow demographic, and generalization to broader populations requires validation across age groups, smoking histories, and device types. Data collection occurred in semi-controlled environments; free-living settings introduce additional confounders (e.g., diverse activities, irregular device wear) that may affect transfer quality. The optimal strategy for collecting personalization data remains unclear: should users label recent events, diverse contexts, or algorithmically selected examples? Active learning approaches that strategically query informative labels could further reduce data requirements[11], an important direction for future work. Additionally, our study focused on a single CNN architecture; investigating transfer across model families (e.g., transformers, recurrent networks) may reveal architecture-dependent benefits.

Future research should also examine the temporal dynamics of personalization. Our study used static train/test splits, but in deployment, models would continuously adapt as users provide labels over time. Online learning strategies that incrementally update models with streaming data could maintain accuracy as individual behaviors evolve (e.g., changes in smoking patterns during cessation attempts). Similarly, exploring meta-learning approaches that optimize for rapid adaptation across individuals[12] may further improve data efficiency.

In conclusion, transfer learning provides a practical pathway to personalized wearable smoking detection, achieving strong performance with minimal individual data collection. By leveraging population knowledge to accelerate individual adaptation, this approach balances accuracy with feasibility, enabling scalable deployment of personalized behavioral interventions. As wearable sensors become ubiquitous in health monitoring, data-efficient personalization strategies will be essential for translating population-level models into individualized precision health tools.

4 Methods

4.1 Participants and Data Collection

Seven participants were recruited for this study [IRB details to be added]. Participants wore wrist-mounted inertial measurement units (IMUs) capturing 6-channel data: 3-axis accelerometer (accel_x, accel_y, accel_z) and 3-axis gyroscope (gyro_x, gyro_y, gyro_z) sampled at 50 Hz. Data collection occurred during [controlled/naturalistic] smoking sessions, with participants instructed to [protocol details]. Smoking bouts were annotated using [annotation method: self-report EMA, observer coding, etc.], with each bout recording start and end timestamps.

The dataset comprised [X total sessions] across seven participants (participant IDs: tonmoy, asfik, alsaad, anam, ejaz, iftakhar, unk1), yielding [X total hours] of sensor data with [Y] annotated smoking bouts. Participants exhibited diverse smoking patterns, hand dominance, and device placement, providing heterogeneity necessary for evaluating personalization.

4.2 Data Preprocessing and Windowing

Raw IMU data were preprocessed using a configuration-driven pipeline (configs/dataset_config.yaml). Sensor streams were segmented into 60-second non-overlapping windows (window_size = 3000 samples at 50 Hz, stride = 3000 samples). Each window was labeled as smoking (positive class) if it overlapped temporally with an annotated smoking bout, otherwise non-smoking (negative class).

Critically, train/validation/test splitting was performed at the session level prior to windowing to prevent temporal data leakage. For each participant, sessions were randomly split (random_state=42) into training (60%), validation (20%), and test (20%) sets. Windows were then extracted independently from each split. This session-level splitting ensures that consecutive windows from the same smoking episode do not appear across train/test boundaries, providing conservative performance estimates.

Final windowed datasets were saved as PyTorch tensors (participant_train.pt, participant_val.pt, participant_test.pt) with shape [N, 6, 3000] for inputs and [N] for binary labels, where N is the number of windows for each split. Class balance varied across participants: [provide statistics on positive class percentage per participant].

4.3 Model Architecture

We employed a convolutional neural network (CNN) architecture designed for time-series classification (TestModel class in lib/models.py). The model consists of:

- **Input layer**: 6 input channels (accelerometer + gyroscope), 3000 timepoints
- **Convolutional blocks**: Four 1D convolutional layers with layer normalization and ReLU activations
  - Block 1: Conv1D(6→8 channels, kernel=3, padding=1) + MaxPool(2)
  - Block 2: Conv1D(8→8, kernel=3, padding=1) + MaxPool(2)
  - Block 3: Conv1D(8→8, kernel=3, padding=1), no pooling
  - Block 4: Conv1D(8→16, kernel=3, padding=1), no pooling
- **Global average pooling**: Adaptive average pooling to reduce temporal dimension to 1
- **Fully connected layer**: Linear(16→1) with sigmoid output for binary classification

The total model contains [X] trainable parameters. Layer normalization was applied to convolutional outputs to stabilize training across participants with different signal amplitudes. The architecture was intentionally kept lightweight to enable rapid fine-tuning and deployment on resource-constrained devices.

4.4 Training Procedures

We compared three training paradigms within a leave-one-participant-out (LOPO) cross-validation framework:

Base Model Training: For each fold, a base model was trained on N-1 participants (N=7, so 6 participants per base model). Training data combined all train and validation splits from base participants (using ConcatDataset). Validation was performed on the combined test splits from base participants. Models were trained using:

- Loss: Binary cross-entropy with logits (BCEWithLogitsLoss)
- Optimizer: AdamW with learning rate 3e-4
- Batch size: [32 or 64, from grid search]
- Early stopping: Patience of 50 epochs based on validation loss

Fine-Tuning: Base model weights were loaded as initialization, and training continued using the target participant's training data. Two fine-tuning variants were evaluated:

- *Full fine-tuning*: Training data = base training data + target training data (combined)
- *Target-only fine-tuning*: Training data = target training data only (no base data in second phase)

Validation used target participant validation split, with early stopping patience of 50 epochs.

Target-Only Training: Models were initialized randomly and trained exclusively on target participant training data, using identical hyperparameters as fine-tuning for fair comparison.

To assess data efficiency, we systematically subsampled target training data to 1%, 5%, 12.5%, 25%, 50%, and 100% of available samples using random_subsample with fixed random seeds. All models were evaluated on the full held-out target participant test set.

Data augmentation was applied during training using TimeSeriesAugmenter with jitter noise (std=0.02), magnitude scaling (range=0.95-1.05), and augmentation probability 0.5 to improve generalization.

4.5 Experimental Design and Hyperparameters

Grid search was performed over:

- Batch size: {32, 64}
- Learning rate: {3e-4}
- Target data percentage: {0.01, 0.05, 0.125, 0.25, 0.5, 1.0}
- Number of base participants (ablation): {1, 2, 3, 4, 5, 6}
- Training mode: {full_fine_tuning, target_only_fine_tuning, target_only}

For each hyperparameter configuration, all seven LOPO folds were executed (generate_jobs.py), yielding [X total training runs]. Experiments were managed using timestamped directories (experiments/{prefix}/fold{N}_{participant with metrics, losses, and model checkpoints saved for reproducibility.

4.6 Evaluation Metrics

Primary evaluation metric was F1 score (harmonic mean of precision and recall), appropriate for the class-imbalanced smoking detection task. Additional metrics included:

- **Precision**: True positives / (true positives + false positives)
- **Recall**: True positives / (true positives + false negatives)
- **Absolute improvement**: F1_personalized - F1_base

- **Relative improvement**: (F1_personalized - F1_base) / F1_base
- **Room for improvement**: (F1_personalized - F1_base) / (1 - F1_base)

All metrics were computed on held-out test sets. Statistical comparisons between training paradigms used [paired t-tests / Wilcoxon signed-rank tests] across the seven LOPO folds, with significance threshold $p < 0.05$. Results are reported as mean ± standard deviation across participants unless otherwise noted.

5 Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Due to privacy considerations regarding participant behavioral data, datasets are not publicly available but can be accessed under data use agreements for research purposes.

6 Code Availability

Code for dataset generation, model training, and evaluation is available at [GitHub repository URL to be added]. The repository includes configuration files (configs/dataset_config.yaml), dataset creation scripts (make_dataset.py), training procedures (train.py, lib/train_utils.py), model architectures (lib/models.py), and evaluation notebooks (eval.py) to ensure full reproducibility of results.

7 Acknowledgements

[Funding sources, technical support, participant acknowledgment to be added]

8 Author Contributions

A.S. designed and implemented the experimental framework, conducted data analysis, and drafted the manuscript. H.V. supervised the research, provided critical feedback on methodology and interpretation, and edited the manuscript. All authors reviewed and approved the final manuscript.

9 Competing Interests

The authors declare no competing interests.

References

[1] World Health Organization. Tobacco fact sheet, 2021.

[2] I. Nahum-Shani, S. N. Smith, B. J. Spring, et al. Just-in-time adaptive interventions (jitais) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

[3] E. Sazonov, P. Lopez-Meyer, and S. Tiffany. A wearable sensor system for monitoring cigarette smoking. *Journal of Studies on Alcohol and Drugs*, 74(6):956–964, 2013.

[4] N. Saleheen, A. A. Ali, S. M. Hossain, et al. puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 999–1010, 2015.

[5] Q. Tang, D. Byrne, I. Nahum-Shani, et al. Toward individualized just-in-time adaptive interventions. *IEEE Journal of Biomedical and Health Informatics*, 24(9):2487–2498, 2020.

[6] A. A. Stone and S. Shiffman. Ecological momentary assessment in behavioral medicine. *Annals of Behavioral Medicine*, 16(3):199–202, 2007.

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[9] A. Parate, M. C. Chiu, C. Chadowitz, et al. Risq: recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th ACM International Conference on Mobile Systems, Applications, and Services*, pages 149–161, 2014.

[10] A. Mathur, M. Van den Broeck, G. Vandenberg, et al. Designing and evaluating contextualized privacy mechanisms in quantified self technologies. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–33, 2021.

[11] B. Settles. Active learning literature survey. *Computer Sciences Technical Report*, 1648, 2009.

[12] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135, 2017.