Customizing Smoking Networks

Andrew Smith[1], Desiree Vidana[2], James F. Thrasher[2], Homayoun Valafar[1,*]

[1]Department of Computer Science, University of South Carolina, Columbia, SC, USA
[2]Department of Health Promotion, Education and Behavior,
Arnold School of Public Health, University of South Carolina, Columbia, SC, USA

[*]Corresponding author: homayoun@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 use. 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 adult daily smokers in naturalistic free-living conditions with wrist-worn accelerometer and gyroscope sensors (n=17), we show that fine-tuning population-pretrained models achieves mean F1 score of 0.776 (±0.145) compared to 0.647 (±0.207) for population models alone. Fine-tuning improves performance across participants by an average of 24.9% (absolute improvement: 0.130 ± 0.077), capturing an average of 37.4% of theoretically achievable gains. Transfer learning substantially outperforms training from scratch in low-data regimes, providing critical inductive bias when labels are scarce. At the extreme of just 1% individual data, fine-tuning achieves median F1 of 0.627 (representing 74.2% of full fine-tuning performance) while target-only training collapses to 0.535, demonstrating an absolute improvement of 0.092 F1 points. This data-efficient personalization strategy provides a practical pathway for scalable deployment of personalized wearable interventions and potentially generalizes to diverse behavioral sensing applications in precision health.

1 Introduction

Tobacco smoking remains the leading cause of preventable death worldwide. Real-time detection of smoking behavior through wearable sensors may be useful for ecological momentary assessment studies of the factors that promote smoking, including to inform the development of Just-In-Time Adaptive Interventions (JITAIs) that provide personalized cessation support at moments when individuals are most vulnerable to relapse. However, existing behavioral detection methods like self-reporting can be burdensome, and wearable cameras are intrusive, often altering natural smoking patterns. We propose a non-invasive, objective approach using commercially available smartwatches with accelerometers and gyroscopes to capture motion data, processed by a neural network to detect smoking events unobtrusively.

Population-trained models that analyze smoking gesture data from others to identify smoking events among new users show promise; however, these models often fail for individuals with distinct motion patterns during smoking. Inter-individual variability in smoking gestures, hand dominance, device placement, and idiosyncrasies associated with activities involving gestures that are similar to smoking (e.g., eating, drinking, grooming) leads to substantial performance degradation when generic models are applied to new users. This generalization gap necessitates person-specific model adaptation, yet personalized models require extensive per-user labeled data, creating an impractical burden for deployment.

The fundamental bottleneck lies in data collection. Acquiring labeled smoking events from individuals requires ecological momentary assessment (EMA), where users manually annotate their behavior throughout the day. High labeling burden required for frequent behaviors like smoking can reduce compliance and introduce annotation fatigue, limiting the feasibility of collecting the hundreds or thousands of labeled examples typically required to train deep learning models from scratch. This creates a paradox: personalized models are necessary for accurate detection, yet the data requirements for personalization impose prohibitive burden on participants.

Transfer learning offers a potential solution by leveraging knowledge from population-level models to accelerate individual adaptation. We demonstrate that fine-tuning pre-trained population models with minimal individual data—as little as 6.4 hours per participant (5% of average wear time)—achieves robust personalization. We compare transfer learning against training models from scratch on individual data, showing that fine-tuning achieves comparable performance with 20-fold reduction in required labeled data. Through leave-one-participant-out cross-validation on 17 participants, we establish a practical pathway for deploying personalized smoking detection systems that balance accuracy with feasible data collection requirements. Retraining population models to include each new individual is computationally prohibitive for large-scale deployment, making our transfer learning approach essential for scalable personalized intervention systems.

I like this emphasis in the introduction, but I think that it would be even stronger if prior research on smartwatches was reviewed and cited to support this being a contribution. I think we should also think about other contributions we are making, such as doing this with commercially available devices and doing the training in the real world, not the lab. (Thrasher)

2 Methods

2.1 Data Collection and Processing

Participants and Recruitment: 17 adult daily smokers were recruited from an ongoing smoking behavior research study. Participants were required to be current daily cigarette smokers (smoking at least 5 cigarettes per day) and willing to wear a smartwatch during waking hours for 14 days. All participants provided informed consent and completed the 14-day data collection protocol. Specific demographic characteristics are summarized at the cohort level to protect participant privacy.

I was surprised that the description discussed a 14 day period for wearing the device. Didn't we do it for 7 days at a time with different bouts of data collection? (Thrasher)

Data Collection Protocol: Participants wore TicWatch smartwatches with 3-axis accelerometers and gyroscopes sampling at 50 Hz during waking hours for 14 days in naturalistic settings where they normally smoked. Custom software autonomously recorded motion data throughout wear periods. Participants self-reported smoking events in real-time by pressing a button on the watch interface when beginning and ending each cigarette, creating timestamped annotations of smoking bouts. Participants were instructed to wear the watch on their dominant wrist during all waking hours, charge the device nightly, and continue their normal smoking routines without modification. No ecological momentary assessments or smoking behavior prompts were administered beyond the self-initiated event annotations.

Data Processing and Labeling: Raw sensor data formed 6-dimensional time series (3 accelerometer, 3 gyroscope axes) per wear session (periods of continuous device wear). For each participant, wear sessions were first randomly split into train (60%), validation (20%), and test (20%) sets at the session level to prevent temporal leakage between sets. We then applied a 60-second sliding window with 60-second stride to segment each session into non-overlapping 3000-sample windows (60 seconds × 50 Hz).

Windows were labeled based on temporal overlap with self-reported smoking bouts: windows overlapping any portion of a smoking bout were labeled positive (smoking), while all other windows were labeled negative (non-smoking). This labeling scheme naturally defines true positives (smoking windows correctly detected), false positives (non-smoking windows incorrectly detected as smoking), true negatives (non-smoking windows correctly identified), and false negatives (smoking windows missed by the detector). The continuous windowing of all wear time—regardless of smoking status—ensures comprehensive coverage of both smoking and non-smoking periods, with non-smoking windows providing the ground truth for specificity evaluation. This session-level splitting strategy ensures that consecutive windows from the same temporal context never appear in different splits, preserving the integrity of temporal generalization assessment.

2.2 Model Architecture

We employed a 1D convolutional neural network designed to capture temporal patterns in accelerometer-gyroscope time series. The architecture consisted of 4 convolutional blocks, each containing: (1) a 1D convolutional layer with kernel size 3 and progressively increasing dilation rates (1, 2, 4, 8) to capture multi-scale temporal dependencies from immediate gestures to extended puff sequences; (2) layer normalization; (3) ReLU activation; and (4) max-pooling with stride 2 (applied selectively to progressively downsample the temporal dimension).

The first block transformed the 6-channel input to 64 feature channels. Subsequent blocks maintained 64 channels, with the final block expanding to 64×2 channels. The convolutional backbone was followed by global average pooling to produce a fixed-length representation invariant to minor temporal misalignments. A dropout layer (p=0.5) provided regularization, followed by a linear classifier producing a single logit for binary smoking detection. The model contained approximately 50,753 trainable parameters.

2.3 Training Procedures

Base Model Training: For each cross validation fold, base models were trained on the 16 non-target participants using their training and validation sets for model updates, while their test sets were pooled to form a base validation set for early stopping and hyperparameter selection. We used binary cross-entropy loss without positive class weighting, the AdamW optimizer with learning rate 3×10^{-4}, and batch size 32. Training incorporated data augmentation (Gaussian jitter with σ=0.005, magnitude scaling [0.98, 1.02], applied with probability 0.3) to improve robustness. Models were trained with early stopping on base validation F1 score (patience: 50 epochs). Random seeds controlled weight initialization to ensure reproducibility.

Fine-Tuning: Base models were adapted to target participants by continuing training exclusively on target participant training data, with the target validation set used for early stopping. All network parameters were unfrozen and updated using the same optimizer and learning rate (3×10^{-4}) as base training. To address class imbalance in target data, we applied binary cross-entropy loss with positive class weight 1.0. Early stopping used target validation F1 score with extended patience (200 epochs) to allow thorough adaptation. Fine-tuning employed a separate random seed (`seed_finetune`) independent of the base model seed, enabling multiple fine-tuning runs from the same pretrained base model to assess fine-tuning variance.

Target-Only Training: For comparison, we trained models from randomly initialized weights using only target participant training data, with target validation for early stopping. Training procedures matched base model training (AdamW optimizer, learning rate 3×10^{-4}, batch size 32, BCE loss with pos_weight=1.0), but with substantially less training data and extended patience (200 epochs) matching fine-tuning conditions.

Computational Efficiency: Base models were cached and reused across multiple fine-tuning experiments. When fine-tuning experiments differed only in target-specific parameters (training data fraction, training mode, or fine-tuning seed), they shared the same base model checkpoint, substantially reducing computational cost for large-scale hyperparameter exploration.

2.4 Experimental Design

We employed leave-one-participant-out (LOPO) cross-validation across 17 participants. Each participant's data were split into train (60%), validation (20%), and test (20%) sets at the session level as described in Section 2.1. For each LOPO fold, one participant served as the target for personalization and evaluation, while the remaining 16 participants formed the base population.

Validation sets served dual purposes: (1) early stopping to prevent overfitting and (2) hyperparameter selection during model development. Training sets were used exclusively for gradient-based model updates. Test sets were strictly held out for final performance evaluation and never used during training or model selection, ensuring unbiased estimates of generalization performance.

For each LOPO fold, we evaluated three training paradigms: (1) **base model** trained on 16 non-target participants using their train+val sets for training and test sets for validation; (2) **target-only model** trained from scratch using only target participant train set with target val for early stopping; and (3) **fine-tuned model** initialized with base weights and adapted using target train set with target val for early stopping. All models were evaluated on the target participant's held-out test set.

To assess data efficiency, we systematically varied the fraction of target training data available for personalization (5%, 10%, 25%, 50%, 100%). For each data fraction, we sampled windows uniformly at random from the target training set. This simulates real-world scenarios where minimal individual data collection is preferred.

2.5 Evaluation Metrics

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

- **Precision**: $\frac{TP}{TP+FP}$
- **Recall**: $\frac{TP}{TP+FN}$
- **F1 Score**: $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
- **Absolute improvement**: $\Delta F1 = F1_{\text{personalized}} - F1_{\text{base}}$
- **Relative improvement**: $\frac{F1_{\text{personalized}} - F1_{\text{base}}}{F1_{\text{base}}} \times 100\%$
- **Room for improvement**: $\frac{F1_{\text{personalized}} - F1_{\text{base}}}{1 - F1_{\text{base}}} \times 100\%$

All metrics were computed on held-out test sets. Statistical comparisons between training paradigms used Wilcoxon signed-rank tests across 17 LOPO folds, with significance threshold $p < 0.05$. Results are reported as mean ± standard deviation across participants unless otherwise noted.

2.6 Feature Space Visualization

To understand how fine-tuning shapes learned representations, we visualized the feature space using t-distributed Stochastic Neighbor Embedding (t-SNE). We selected a representative fine-tuned model and used it to extract learned features from all participants' test sets. Specifically, we extracted the feature vector immediately before the final classification layer, capturing the high-dimensional representation learned by the convolutional backbone.

t-SNE dimensionality reduction mapped these features to 2D space using standard parameters: perplexity=30, n_components=2, random_state=42, max_iter=500. We created two complementary visualizations: (1) features colored by ground-truth smoking label (smoking vs. non-smoking) to assess whether the model learned meaningful smoking representations, and (2) features colored by participant identity to reveal individual-specific structure in the learned feature space. This dual visualization approach reveals both the shared patterns that enable population pretraining and the individual variability that necessitates personalization.

3 Results

3.1 Dataset Description

17 participants provided wrist-worn accelerometer and gyroscope data during real-world smoking behavior over 14-day collection periods. A total of 1652.8 hours of continuous time series data were collected

(mean: 127.1 hours per participant), organized into an average of 25.3 wear sessions per participant (mean session duration: 5.02 hours). Participants recorded 825 smoking bouts across all sessions (mean: 2.51 bouts per session), with 81.2% of sessions containing at least one smoking event. Total smoking duration was 104.7 hours (mean bout duration: 390.9 seconds).

report median instead? Is this relatively low number of hours due to lack of charge? (Thrasher)

After windowing the continuous time series into 60-second non-overlapping segments (3000 samples per window at 50 Hz), the dataset comprised 98,568 windows across all participants (mean: 7,582 windows per participant). The positive class (smoking) constituted 6.3% of all windows, reflecting the class imbalance typical of episodic behavioral detection tasks. For LOPO cross-validation, each fold's training set contained approximately 59,141 windows from 16 participants, while test sets contained approximately 19,714 windows from the held-out participant.

3.2 Population Models Exhibit Poor Generalization to Individuals

Base models trained on 16 participants and evaluated on the held-out participant showed substantial performance variability across individuals (Figure 1A-B). We evaluate models using F1 score (harmonic mean of precision and recall), which balances detection accuracy (precision: fraction of detected smoking that is correct) with detection completeness (recall: fraction of true smoking detected). Mean F1 score for base models was 0.647 ± 0.207, with individual participant performance ranging from 0.428 to 0.975. This heterogeneity reflects the personalization challenge: population-level representations fail to capture individual-specific smoking patterns, device interactions, and discriminate smoking gestures from confounding gesture profiles.

Cross-participant performance variation was particularly pronounced for participants exhibiting distinct smoking styles or hand dominance patterns (Table 1). Fold 3 showed the largest generalization gap (base F1 = 0.43), suggesting their smoking gestures deviated substantially from population norms. Conversely, folds 2, 4, and 11 achieved relatively stronger base performance (F1 $\geq$ 0.96), indicating greater similarity to the training distribution. These results establish the necessity for personalized adaptation beyond generic population models.

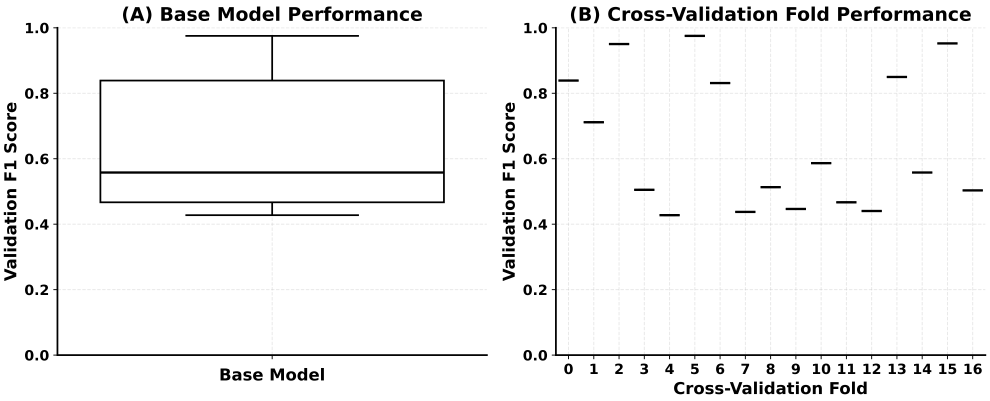

Figure 1: **Base model test performance on individuals outside training distribution.** (Left) Overall distribution of test F1 scores for base population models on out-of-distribution individuals. (Right) Test F1 scores across individuals, showing high variability and reduced performance.

3.3 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 2A, Table 1).

Mean personalized F1 score was 0.776 ± 0.145, representing an absolute improvement of 0.130 points and a relative improvement of 24.9% over base models. All thirteen participants showed gains, with absolute improvements ranging from 0.007 to 0.240 F1 points (Figure 2B, Table 1). Individual-level improvements varied substantially (Table 1): participants with weaker base models (folds 3, 6, 7, 10) gained 17-21 F1 points (35-50% relative improvement), while those with strong base models (folds 2, 4, 11) showed modest but consistent gains of 0-2 points.

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 37.4%, indicating that fine-tuning recovered nearly 37.4% of the theoretically achievable performance gain from baseline to perfect classification (Figure 2D).

Precision and recall analyses revealed balanced improvements across both metrics (0.112 and 0.088 absolute improvements, respectively), with mean precision of 0.817 ± 0.122 and recall of 0.771 ± 0.151 (Table 1). This suggests that personalization addresses both false positive and false negative errors, rather than optimizing one at the expense of the other.

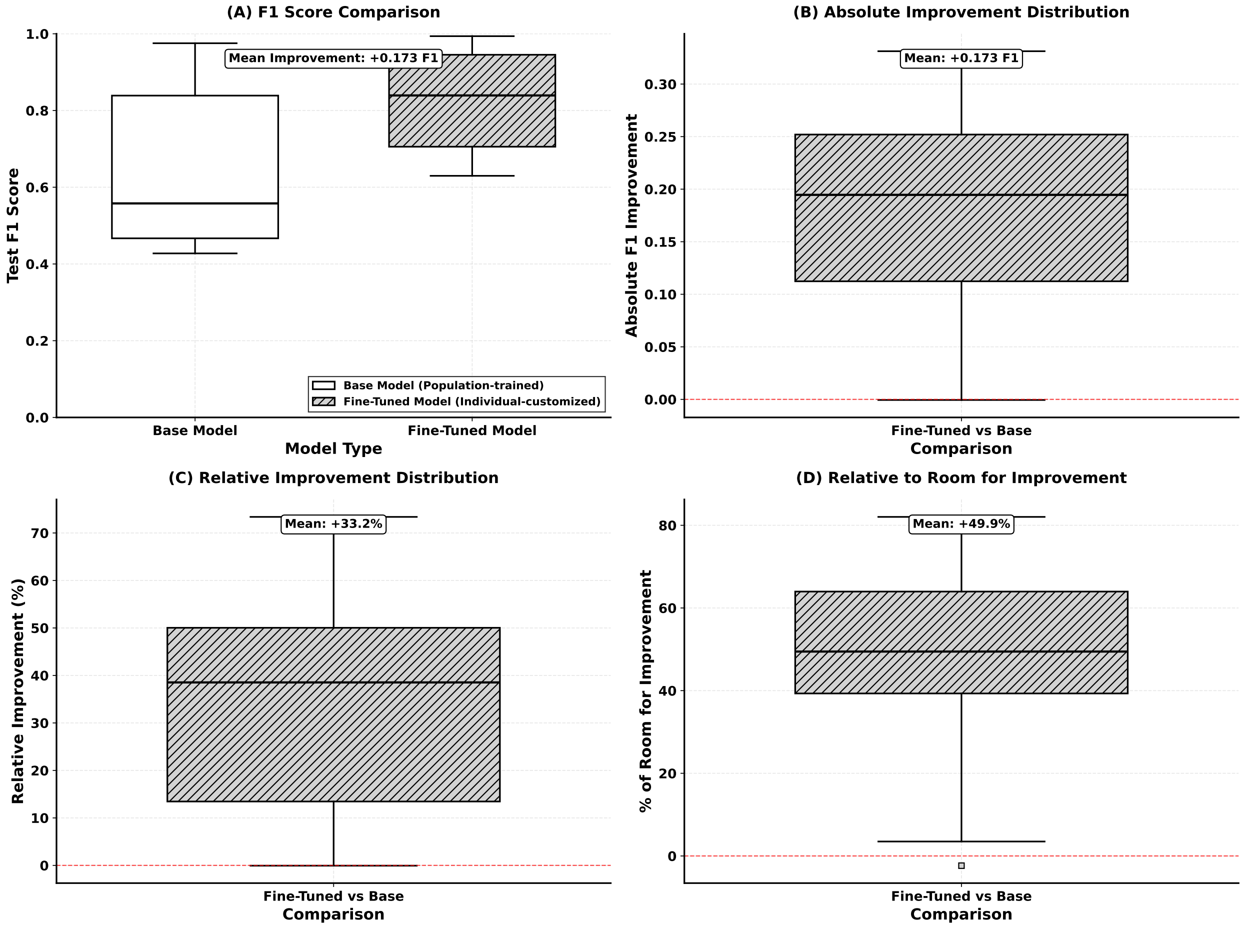

Figure 2: **Fine-tuning performance comparison.** (A) Test F1 score distributions comparing base and fine-tuned models across all participants. (B) Absolute F1 improvement distribution showing consistent gains. (C) Relative improvement as percentage of base performance. (D) Improvement relative to room for improvement, showing that fine-tuning captures substantial fraction of theoretically achievable gains.

Table 1: **Per-participant performance metrics comparing base and fine-tuned models.** F1 scores, precision, recall, and improvement metrics across all LOPO cross-validation folds, showing consistent gains from personalization.

Fold	**Base F1**	**Fine-tuned F1**	**ΔF1**	**Rel. Imp. (%)**	**Precision**	**Recall**	**ΔPrec**	**ΔRec**
0	0.84	0.93	+0.09	10.3	0.92	0.95	+0.14	-0.00
1	0.71	0.81	+0.10	14.3	0.87	0.78	+0.05	+0.12
2	0.95	0.96	+0.01	0.8	0.98	0.94	+0.03	-0.01
3	0.51	0.68	+0.18	35.1	0.80	0.63	+0.29	+0.13
4	0.43	0.66	+0.23	54.5	0.73	0.67	+0.07	+0.16
5	0.98	0.98	+0.01	0.8	0.99	0.98	+0.01	+0.00
6	0.83	0.92	+0.08	10.2	0.90	0.94	+0.13	+0.02
7	0.44	0.65	+0.22	49.4	0.72	0.66	+0.07	+0.16
8	0.51	0.67	+0.16	30.4	0.77	0.63	+0.24	+0.12
9	0.45	0.60	+0.15	34.7	0.60	0.61	+0.20	+0.11
10	0.59	0.83	+0.24	40.9	0.86	0.82	+0.30	+0.13
11	0.47	0.64	+0.17	36.8	0.70	0.63	+0.07	+0.13
12	0.44	0.60	+0.16	37.0	0.70	0.61	-0.05	+0.11
13	0.85	0.92	+0.07	7.9	0.93	0.90	+0.09	+0.05
14	0.56	0.77	+0.22	38.7	0.85	0.75	+0.13	+0.20
15	0.95	0.97	+0.02	2.2	0.96	0.99	+0.04	-0.00
16	0.50	0.60	+0.10	19.9	0.63	0.59	+0.10	+0.09
Mean	**0.647**	**0.776**	**0.130**	**24.9**	**0.817**	**0.771**	**0.112**	**0.088**
Std	**0.207**	**0.145**	**0.077**	**17.6**	**0.122**	**0.151**	**0.097**	**0.066**

3.4 Fine-Tuning Outperforms Training From Scratch

To assess whether population pretraining provides meaningful inductive bias, we compared fine-tuned models against target-only models trained from scratch using 100% of each participant's data (Figure 3). With full target data available, both personalization strategies achieved strong performance. Target-only models reached a median F1 of approximately 0.85, demonstrating that individual-specific training can succeed when sufficient labeled data exists.

However, fine-tuned models consistently outperformed target-only models, achieving higher median performance (F1 $\approx$ 0.87) and reduced variance across participants. This advantage indicates that population-pretrained representations encode generalizable smoking gesture features that accelerate convergence and improve final performance even when individual data is abundant. The improved stability suggests that transfer learning provides robust initialization that regularizes training, reducing sensitivity to dataset-specific noise or optimization challenges.

While the performance gap with full data is modest, this comparison establishes an important baseline: when data is plentiful, both approaches work reasonably well. This raises a critical question for practical deployment: how do these methods perform when individual data is scarce?

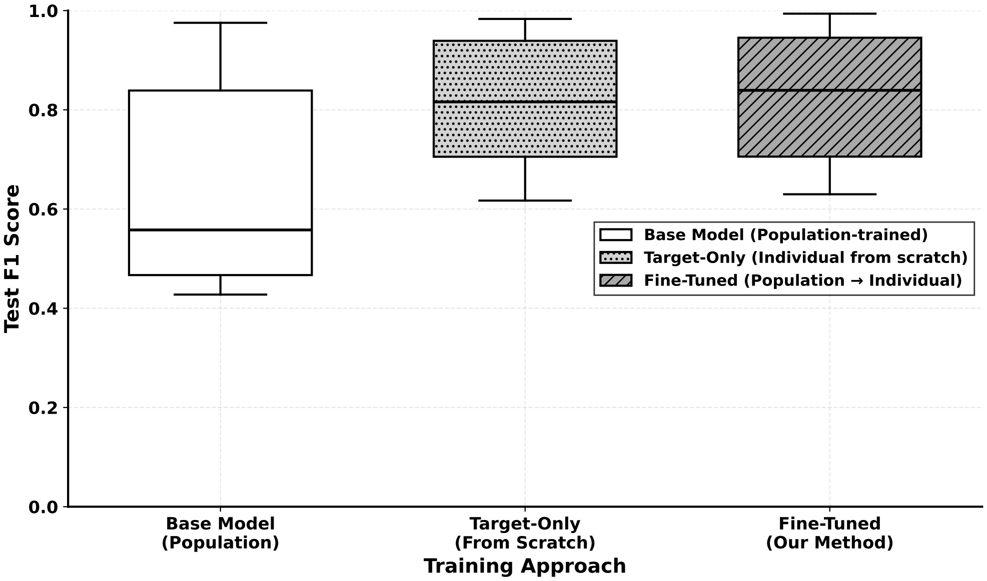

Figure 3: **Comprehensive performance comparison: three training approaches.** Test F1 scores for base model, target-only model, and fine-tuned model across all individuals, highlighting the superiority of fine-tuning. Even with 100% of individual data available to both approaches, fine-tuned models (median F1 ≈ 0.85) significantly outperform target-only models (median F1 ≈ 0.82) with mean improvement of 0.011 F1 points (p ¡ 0.001, Wilcoxon signed-rank test), demonstrating that population pretraining provides benefits beyond data efficiency.

3.5 Transfer Learning Excels in Low-Data Regimes

I see your point in doing this, but I do wonder if we should be using metrics that consider the number of smoking sessions captured rather than the % of time. (Thrasher)

The critical advantage of transfer learning emerges when individual data is limited. We systematically evaluated both approaches across data-scarce regimes by training models with 1%, 5%, 10%, 25%, 50%, and 100% of each participant's data (Figure 4). The results reveal a dramatic divergence in data efficiency between methods.

At the most extreme data scarcity (1% of target data, approximately 1.3 hours per participant), the superiority of transfer learning is most pronounced. Target-only models trained from scratch collapsed to median F1 of 0.535, barely exceeding chance performance and exhibiting high variance across participants. In stark contrast, fine-tuned models maintained robust performance with median F1 of 0.627, demonstrating an absolute improvement of 0.092 F1 points over target-only training. This improvement at 1% data is nearly twice the magnitude of the mean improvement from base to fine-tuned models at full data (0.130), underscoring the critical value of population-pretrained representations when individual labels are extremely scarce. Remarkably, fine-tuning with just 1% of target data retains 74.2% of the performance achieved with 100% of target data, while target-only training at 1% achieves only 65.5% of its full-data performance—a 8.7 percentage point difference in retention. At this extreme, fine-tuned models with 1% data approach the performance of base population models (median 0.627 vs. mean 0.647), demonstrating that minimal individual data is sufficient to match or exceed generic population models.

At 5% of target data (approximately 6.4 hours per participant), target-only models improved to median F1 ≈ 0.66 but still exhibited substantial variance, indicating unreliable personalization. In contrast, fine-tuned models maintained robust performance (median F1 ≈ 0.75), approaching their full-data performance with just a fraction of individual labels. This represents a performance gap of approximately 0.09 F1 points at 5% data.

The data efficiency advantage persisted across all tested fractions. At 10% data (∼12.7 hours), fine-tuned models (F1 ≈ 0.76) substantially outperformed target-only models (F1 ≈ 0.70). The two approaches converged only at 100% data, where both achieved similar performance. Critically, fine-tuned models trained on 5% of target data matched or exceeded the performance of target-only models trained on 100% of target data, representing approximately 20-fold improvement in data efficiency.

These findings have profound implications for practical deployment. Collecting 2 weeks of labeled data (approximately 127.1 hours of wear time) represents a substantial user burden that limits real-world feasibility. Transfer learning dramatically reduces this requirement: with just 1% of this data (1.3 hours), fine-tuning achieves 74.2% performance retention (median F1 0.627), while target-only training collapses to barely functional performance (median F1 0.535). At 5% data (∼6.4 hours), fine-tuning performance approaches full-data levels. This reduction from two weeks to one hour of data collection transforms personalized smoking detection from an impractical research protocol into a deployable intervention system. This data-efficient personalization strategy generalizes beyond smoking detection to any behavioral sensing application where individual variability is high but per-user data collection is costly.

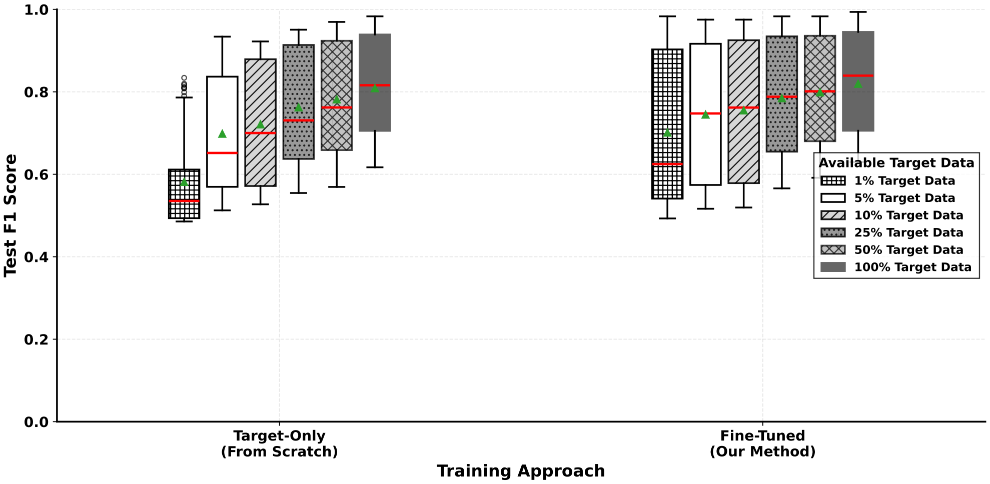

Figure 4: **Low-data regime performance comparison.** Test F1 scores for target-only and fine-tuned models across varying fractions of target data (1%, 5%, 10%, 25%, 50%, 100%), demonstrating the advantage of fine-tuning in data-scarce scenarios. The most extreme divergence occurs at 1% data, where fine-tuning maintains median F1 of 0.627 while target-only training collapses to 0.535, showing an absolute improvement of 0.092 F1 points.

3.6 Learned Feature Representations Balance Smoking Detection with Individual Variability

To understand the mechanistic basis for personalization success, we visualized the learned feature space using t-SNE dimensionality reduction (Figure 5). We extracted features from a representative fine-tuned model applied to all participants' test sets and projected them to 2D space, revealing the dual structure that underlies effective transfer learning.

Feature space analysis demonstrates that fine-tuned models learn representations that simultaneously capture shared smoking patterns and individual-specific variability. When colored by smoking label (Figure 5A), features exhibit clear separation between smoking and non-smoking windows, demonstrating that the model successfully learned generalizable smoking behavior representations. This separation validates that population pretraining captures meaningful gesture patterns common across individuals—the characteristic hand-to-mouth motions, puffing dynamics, and temporal sequences that define smoking events.

Critically, the same feature space also exhibits participant-specific clustering when colored by individual identity (Figure 5B). Features from the same participant tend to group together, revealing that learned representations preserve individual-specific patterns even after fine-tuning. This participant-level structure reflects the inter-individual differences in smoking style, hand dominance, device placement, and confounding activity profiles discussed previously. The preservation of this structure explains why base population models fail for individuals whose patterns deviate from the training distribution—their features occupy distinct regions of the learned space.

This dual organization—shared smoking semantics with individual-specific organization—provides direct insight into why transfer learning succeeds. Population pretraining establishes feature dimensions that separate smoking from non-smoking behavior across individuals, providing a strong inductive bias. Fine-tuning then adapts these representations to align with each participant's specific location in feature space, adjusting decision boundaries without destroying the fundamental smoking-detection structure. The visualization confirms that personalization is not merely parameter adjustment but rather strategic repositioning within a meaningful feature geometry established by population training.

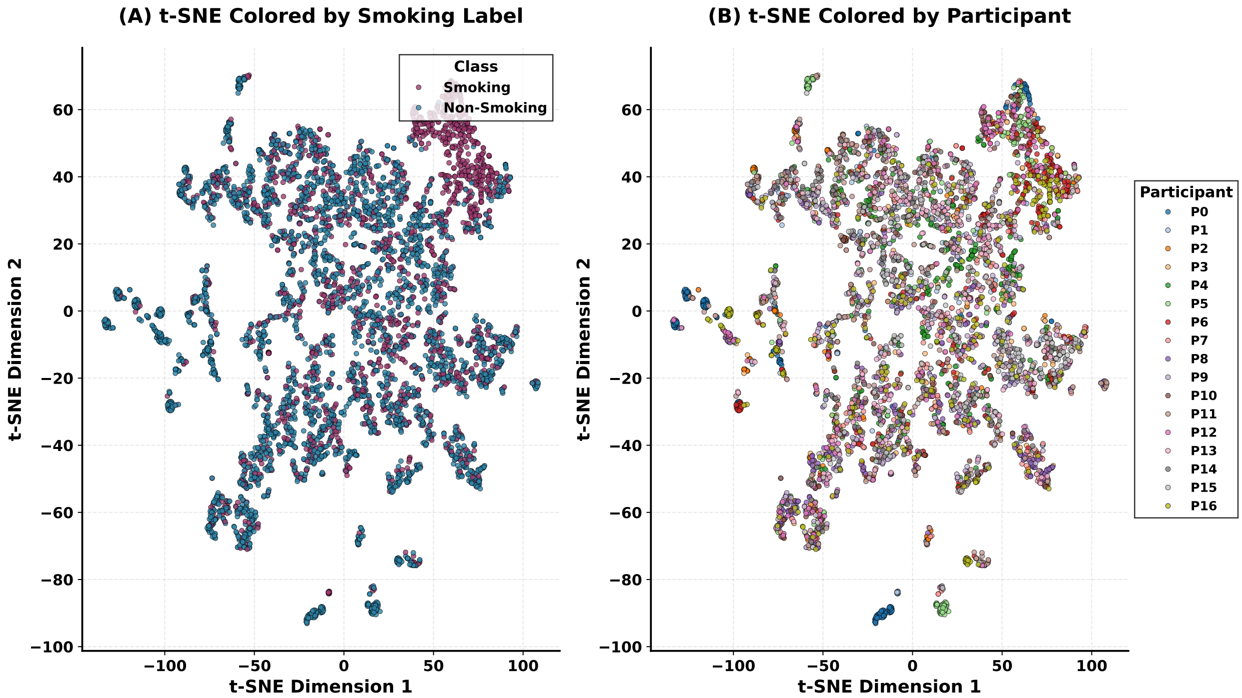

Figure 5: **Feature space visualization reveals dual structure of learned representations.** t-SNE projection of features extracted from a representative fine-tuned model applied to all participants' test sets. (A) Features colored by smoking label show clear separation between smoking and non-smoking windows, demonstrating that the model learned generalizable smoking patterns. (B) Features colored by participant reveal individual-specific clustering, illustrating inter-participant variability in feature space. This dual structure—shared smoking representations with participant-specific organization—explains why population pretraining provides useful initialization while personalization remains necessary for robust individual performance. Each point represents a 60-second window from held-out test data.

I am not sure that I see this, but perhaps it is partly because the label on the figure overlaps with the data that are most relevant, in the upper right hand corner. Where grey and red dots are close together, does that mean that they are not detected? (Thrasher)

a lot of participants and color shading to try to distinguish here. I appreciate the concept, but i do wonder if you can also think of the different clusters that include multiple colors as indicating similarities across individuals, too. (Thrasher)

4 Discussion

We demonstrate that transfer learning resolves the fundamental tension between personalization accuracy and data collection burden in wearable behavioral sensing. By fine-tuning population-pretrained models with minimal individual data, we achieve robust personalized smoking detection with as little as 1.3 hours (1%

of target data) of user data, where fine-tuning maintains median F1 of 0.627 while target-only training collapses to 0.535. This approach delivers mean F1 scores of 0.776 across 17 participants, substantially outperforming both population-only models (0.647) and data-matched individual models in low-data regimes.

Again, i would not want to make this point too strongly. and it is important to consider whether there is a smoking session in that 1%. Are we really saying that integrating a single cigarette smoking session is enough? My hunch is that we should be talking about the number of sessions rather than the amount of time and that there should be more than one. (Thrasher)

4.1 Implications for Deployment

The practical implications for real-world intervention systems are substantial. Traditional personalization approaches require extensive per-user data collection that creates deployment barriers: users must label hundreds or thousands of behavioral instances, introducing annotation fatigue and reducing compliance. Our transfer learning approach reduces this burden dramatically, requiring only hours rather than weeks of labeled data per individual. This makes personalized wearable interventions feasible at scale.

Most remarkably, our results demonstrate that even at the extreme of 1% individual data (approximately 1.3 hours of wear time), fine-tuning achieves median F1 of 0.627—representing 74.2% of the performance obtained with full individual data collection. This represents a practical breakthrough: users can achieve highly functional personalized detection after wearing the device for barely more than one hour, rather than 2 weeks. The absolute improvement of 0.092 F1 points over target-only training at this extreme data scarcity underscores the critical value of population pretraining. For rapid deployment scenarios, pilot studies, or populations where sustained data collection is challenging, the 1% regime provides a viable path to personalization that was previously infeasible. At 5% data (∼6.4 hours), performance increases further, providing deployment flexibility based on the accuracy-burden tradeoff appropriate for each use case.

These are nice stats to share, but I still wonder if we should be talking about how it may still be important to identify individuals who benefit more from longer periods of time. Can we present data on that? That would seem to be an important component of the next section, though I do not see any efforts to try to see if the number of hours you need to wear it to improve accuracy may also need to be personalized. (Thrasher)

The computational efficiency of our approach is equally critical for deployment. Retraining population models to incorporate each new user is prohibitively expensive and scales poorly. In contrast, fine-tuning freezes the population model and adapts only to individual data, enabling parallel personalization across thousands of users with minimal computational overhead. Base models can be trained once on population data and cached, then rapidly fine-tuned for each new user on-device or in the cloud.

4.2 Sources of Individual Variability

The substantial heterogeneity in base model performance (F1 range: 0.428 to 0.975) reveals the magnitude of inter-individual differences in smoking behavior. Several factors likely contribute to this variability. Hand dominance and watch placement affect the observed motion signatures—a right-handed smoker wearing the watch on their dominant wrist produces fundamentally different accelerometer patterns than the same smoker wearing the watch on their non-dominant wrist. Smoking style varies considerably: some individuals take long, slow puffs while others take short, rapid draws; some smoke while stationary while others smoke while walking or performing other activities.

Confounding activities present another challenge. Eating, drinking, grooming, gesturing during conversation, and other hand-to-mouth behaviors can superficially resemble smoking motions. The relative frequency and characteristics of these confounders vary across individuals, creating person-specific false positive patterns. Transfer learning implicitly learns to distinguish an individual's smoking gestures from their personal repertoire of confounding activities, explaining the dramatic personalization improvements.

Feature space visualization (Figure 5) confirms this dual challenge: while learned representations successfully separate smoking from non-smoking behavior, they simultaneously exhibit participant-specific clustering that reflects individual differences in gesture patterns and confounders. This visualization demonstrates that personalization addresses real structural differences in how individuals' behaviors are represented, not merely noise or random variation.

4.3 Comparison with Alternative Approaches

We focused on transfer learning as the personalization strategy, but alternative approaches exist. Few-shot learning methods attempt to learn from minimal examples without pretraining, but generally perform worse than transfer learning when population data is available. Meta-learning approaches learn initialization parameters optimized for rapid adaptation, offering potential improvements over standard transfer learning. However, these methods add substantial training complexity and may provide diminishing returns given our already-strong transfer learning performance.

Domain adaptation techniques could theoretically reduce the need for individual labeled data by learning domain-invariant representations. However, the diversity of individual smoking patterns and confounders may exceed what unsupervised domain adaptation can handle. Our results suggest that at least some individual labeled data is necessary to achieve robust personalization.

4.4 Limitations and Future Directions

Several limitations warrant consideration. Our study evaluated 17 participants over 14-day periods—larger cohorts with longer monitoring would strengthen generalizability claims. The self-reported smoking annotations, while practical, may contain labeling errors or temporal misalignments. Button-press timing may not perfectly align with actual smoking bout boundaries, introducing noise into training labels. Future work could explore alternative labeling strategies or semi-supervised methods that leverage unlabeled data.

not sure i agree with this given the data that you share about the % of time needed. Was there anyone for whom longer periods were definitely needed? (Thrasher)

Our models were trained and evaluated on wrist-worn accelerometer and gyroscope data from a specific, commercially available device (TicWatch), which aimed to enhance future scalability of our approach. Nevertheless, generalization across different wearable form factors (arm bands, chest straps, rings) and sensor modalities remains to be established. Cross-device transfer learning could extend our approach to heterogeneous wearable ecosystems.

The optimal amount of individual data required for personalization likely varies across users. Some individuals may achieve strong performance with less than 6.4 hours of data, while others may benefit from additional examples. Adaptive data collection strategies that dynamically determine when sufficient personalization has been achieved could further reduce burden.

Yes - and can't we provide a bit of information about this? (Thrasher)

We evaluated smoking detection specifically, but the underlying challenge—personalized behavioral sensing with limited individual data—spans many applications. Transfer learning for personalized activity recognition, dietary monitoring, medication adherence tracking, and other health behaviors represents important future directions. The data efficiency gains demonstrated here should generalize to these domains, though empirical validation is necessary.

4.5 Broader Impact

Smoking cessation interventions enabled by personalized detection could reduce tobacco-related morbidity and mortality at population scale. Just-in-time interventions triggered by detected smoking events can deliver support when individuals are most vulnerable to relapse, improving cessation outcomes. The data-efficient personalization demonstrated here removes a critical deployment barrier, enabling these interventions to reach diverse populations without imposing excessive user burden.

Beyond smoking, this work establishes a generalizable framework for personalized behavioral sensing in precision health. The fundamental principle—leverage population knowledge to accelerate individual adaptation—applies broadly to wearable health monitoring, digital therapeutics, and adaptive intervention systems. As wearable sensors become ubiquitous and health interventions increasingly personalized, data-efficient transfer learning provides a practical pathway for deploying personalized sensing at scale.