

Customizing Smoking Networks

Andrew Smith¹, Desiree Vidana², James F. Thrasher², Homayoun Valafar^{1,*}

¹Department of Computer Science, University of South Carolina, Columbia, SC, USA

²Department 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 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 17 participants with wrist-worn accelerometer and gyroscope sensors, 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 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 enables Just-In-Time Adaptive Interventions (JITAIs), providing personalized cessation support at moments when individuals are most vulnerable to relapse. However, existing detection methods like self-reporting or wearable cameras are intrusive, often altering natural smoking patterns. We propose a non-invasive approach using smartwatches with accelerometers and gyroscopes to capture motion data, processed by a neural network to detect smoking events unobtrusively.

While population-trained models show promise, they often fail for individuals with distinct motion patterns during smoking. Inter-individual variability in smoking gestures, hand dominance, device placement, and confounding activities such as eating, drinking, or 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 disruptive 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. 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. We demonstrate that fine-tuning pre-trained population models with minimal

40 individual data—as little as 6.4 hours per participant (5% of average wear time)—achieves robust person-
41 alization. We compare transfer learning against training models from scratch on individual data, showing
42 that fine-tuning achieves comparable performance with 20-fold reduction in required labeled data. Through
43 leave-one-participant-out cross-validation on 17 participants, we establish a practical pathway for deploying
44 personalized smoking detection systems that balance accuracy with feasible data collection requirements.
45 Retraining population models to include each new individual is computationally prohibitive for large-scale
46 deployment, making our transfer learning approach essential for scalable personalized intervention systems.

47 2 Related Works

48 2.1 Smoking Detection and the Personalization Challenge

49 Wearable sensor-based smoking detection has evolved from traditional machine learning to deep neural
50 architectures that capture hand-to-mouth gestures through accelerometer and gyroscope signals. CNN-
51 LSTM hybrid architectures achieve F1 scores around 0.91 for puff detection using respiratory sensors [1], while
52 smartwatch-based ANNs demonstrate 85-95% accuracy in controlled settings [2]. Hierarchical approaches
53 combining neural networks with rule-based filters yield 89-99% true positive rates in free-living scenarios [2],
54 and systematic reviews report accuracies up to 98.5% for deep learning methods on motion and physiological
55 signals [3].

56 However, these population-trained models exhibit substantial performance degradation when deployed to
57 new individuals due to inter-person variability in smoking gestures, hand dominance, device placement, and
58 confounding activities (eating, drinking, grooming). This generalization gap necessitates personalized adap-
59 tation, yet training individual models from scratch requires extensive per-user labeled data—a prohibitive
60 burden for real-world deployment where users must manually annotate smoking events throughout multi-day
61 collection periods.

62 2.2 Transfer Learning for Personalized Activity Recognition

63 Transfer learning addresses this data scarcity challenge by adapting population-pretrained models to individ-
64 uals with minimal target data. Comprehensive surveys categorize approaches into instance transfer (domain
65 alignment via MMD or synthetic data generation), feature transfer (space alignment through TCA or con-
66 trastive learning), and parameter transfer (pretrain-then-finetune paradigms) [4]. Recent HAR personaliza-
67 tion frameworks predominantly employ unsupervised or few-shot methods to minimize labeling requirements:
68 unsupervised domain adaptation using JPDA with pseudo-label refinement achieves 93.2% accuracy across
69 activities [5], Bayesian networks with active learning reach 89.2% accuracy by selectively querying uncertain
70 samples [6], and teacher-student self-training with contrastive learning enables few-shot cross-domain adap-
71 tation [7]. Federated learning variants allow privacy-preserving personalization by fine-tuning global models
72 on local data [8].

73 While these methods reduce labeling burden, they face limitations for smoking detection’s specific chal-
74 lenges. Unsupervised approaches risk error propagation from inaccurate pseudo-labels when distinguishing
75 subtle smoking gestures from similar confounding activities. Few-shot methods may lack sufficient signal to
76 capture individual-specific patterns like unique puff durations or non-dominant hand use. Critically, existing
77 work provides limited systematic comparison of how transfer learning performs against training from scratch
78 across varying data regimes—a key practical question for deployment planning.

79 2.3 Contribution and Positioning

80 Our work addresses these gaps through three contributions. First, we apply supervised fine-tuning of
81 population-pretrained models specifically to smoking detection, leveraging explicit individual labels to achieve
82 precise adaptation to person-specific gesture patterns and confounders. Second, we provide systematic quan-
83 titative comparison between fine-tuning and training from scratch across extreme data scarcity regimes (1%,
84 5%, 10%, 25%, 50%, 100% of individual data), revealing that fine-tuning with just 1% target data (approx-
85 imately 1.3 hours) achieves median F1 of 0.627 while target-only training collapses to 0.535—an absolute
86 improvement of 0.092 F1 points. Third, we demonstrate approximately 20-fold data efficiency improvement,

87 where fine-tuning with 5% individual data matches target-only performance at 100% data. These findings
88 establish a practical deployment pathway for personalized smoking interventions and generalize to behavioral
89 sensing applications where individual variability is high but per-user data collection is costly.

90 3 Methods

91 3.1 Data Collection and Processing

92 **Participants and Recruitment:** [TODO: Replace with actual recruitment details] 17 adult daily smokers
93 were recruited from an ongoing smoking behavior research study. Participants were required to be current
94 daily cigarette smokers (smoking at least 5 cigarettes per day) and willing to wear a smartwatch continuously
95 for 14 days. All participants provided informed consent and completed the 14-day data collection protocol.
96 Specific demographic characteristics are summarized at the cohort level to protect participant privacy.

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

105 **Data Processing and Labeling:** Raw sensor data formed 6-dimensional time series (3 accelerometer,
106 3 gyroscope axes) per wear session (periods of continuous device wear). For each participant, wear sessions
107 were first randomly split into train (60%), validation (20%), and test (20%) sets at the session level to
108 prevent temporal leakage between sets. We then applied a 60-second sliding window with 60-second stride
109 to segment each session into non-overlapping 3000-sample windows (60 seconds \times 50 Hz).

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

119 3.2 Model Architecture

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

126 The first block transformed the 6-channel input to 64 feature channels. Subsequent blocks maintained
127 64 channels, with the final block expanding to 64×2 channels. The convolutional backbone was followed by
128 global average pooling to produce a fixed-length representation invariant to minor temporal misalignments.
129 A dropout layer ($p=0.5$) provided regularization, followed by a linear classifier producing a single logit for
130 binary smoking detection. The model contained approximately 50,753 trainable parameters.

131 3.3 Training Procedures

132 **Base Model Training:** For each LOPO fold, base models were trained on the 16 non-target participants
133 using their training and validation sets for model updates, while their test sets were pooled to form a base

134 validation set for early stopping and hyperparameter selection. We used binary cross-entropy loss without
135 positive class weighting, the AdamW optimizer with learning rate 3×10^{-4} , and batch size 32. Training
136 incorporated data augmentation (Gaussian jitter with $\sigma=0.005$, magnitude scaling [0.98, 1.02], applied with
137 probability 0.3) to improve robustness. Models were trained with early stopping on base validation F1 score
138 (patience: 50 epochs). Random seeds controlled weight initialization to ensure reproducibility.

139 **Fine-Tuning:** Base models were adapted to target participants by continuing training exclusively on
140 target participant training data, with the target validation set used for early stopping. All network param-
141 eters were unfrozen and updated using the same optimizer and learning rate (3×10^{-4}) as base training.
142 To address class imbalance in target data, we applied binary cross-entropy loss with positive class weight
143 1.0. Early stopping used target validation F1 score with extended patience (200 epochs) to allow thorough
144 adaptation. Fine-tuning employed a separate random seed (`seed_finetune`) independent of the base model
145 seed, enabling multiple fine-tuning runs from the same pretrained base model to assess fine-tuning variance.

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

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

154 3.4 Experimental Design

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

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

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

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

172 3.5 Evaluation Metrics

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

175 • **Precision:** $\frac{TP}{TP+FP}$

176 • **Recall:** $\frac{TP}{TP+FN}$

177 • **F1 Score:** $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

178 • **Absolute improvement:** $\Delta F1 = F1_{\text{personalized}} - F1_{\text{base}}$

179 • **Relative improvement:** $\frac{F1_{\text{personalized}} - F1_{\text{base}}}{F1_{\text{base}}} \times 100\%$

180 • Room for improvement: $\frac{F1_{\text{personalized}} - F1_{\text{base}}}{1 - F1_{\text{base}}} \times 100\%$

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

184 3.6 Feature Space Visualization

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

190 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:
 191 (1) features colored by ground-truth smoking label (smoking vs. non-smoking) to assess whether
 192 the model learned meaningful smoking representations, and (2) features colored by participant identity to
 193 reveal individual-specific structure in the learned feature space. This dual visualization approach reveals
 194 both the shared patterns that enable population pretraining and the individual variability that necessitates
 195 personalization.

197 4 Results

198 4.1 Dataset Description

199 17 participants provided wrist-worn accelerometer and gyroscope data during real-world smoking behavior
 200 over 14-day collection periods. A total of 1652.8 hours of continuous time series data were collected (mean:
 201 127.1 hours per participant), organized into an average of 25.3 wear sessions per participant (mean session
 202 duration: 5.02 hours). Participants recorded 825 smoking bouts across all sessions (mean: 2.51 bouts per
 203 session), with 81.2% of sessions containing at least one smoking event. Total smoking duration was 104.7
 204 hours (mean bout duration: 390.9 seconds).

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

211 4.2 Population Models Exhibit Poor Generalization to Individuals

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

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

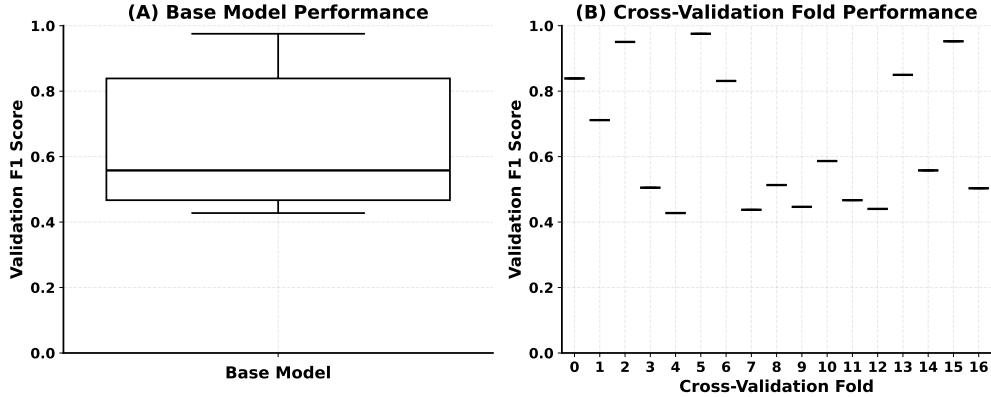


Figure 1: **Base model test performance on individuals outside training distribution.** (Left) Overall distribution of test F1 scores for base population models on out-of-distribution individuals. (Right) Test F1 scores across individuals, showing high variability and reduced performance.

225 4.3 Fine-Tuning Achieves Robust Personalization with Full Target Data

226 Fine-tuning base models with 100% of target participant training data yielded substantial performance
227 improvements across all participants (Figure 2A, Table 1).

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

234 We quantified improvement using three complementary metrics: (1) **absolute improvement** (per-
235 sonalized F1 - base F1); (2) **relative improvement** ((personalized - base) / base); and (3) **room for
236 improvement** (absolute improvement / (1 - base F1)), which accounts for each participant's potential for
237 gains. Mean room-for-improvement capture was 37.4%, indicating that fine-tuning recovered nearly 37.4%
238 of the theoretically achievable performance gain from baseline to perfect classification (Figure 2D).

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

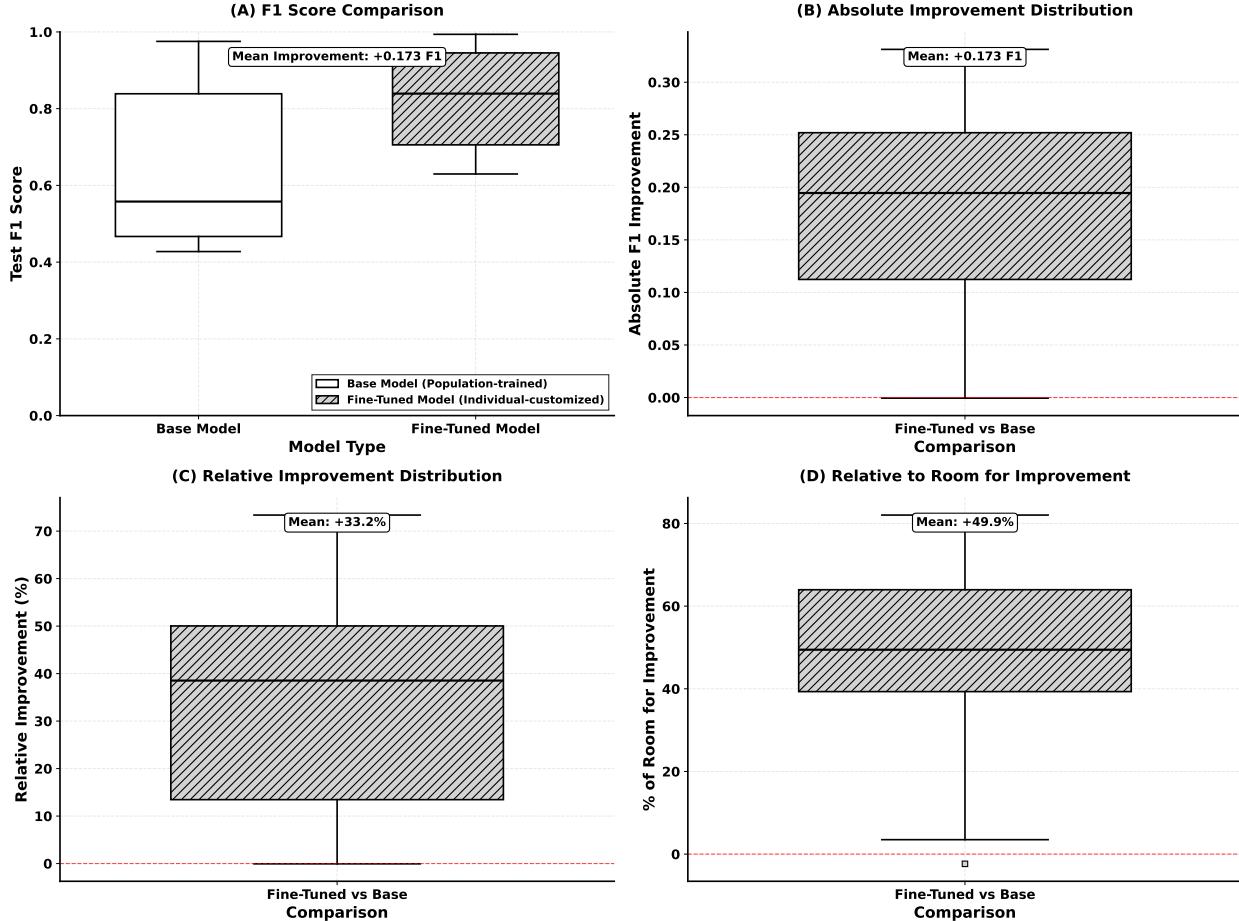


Figure 2: **Fine-tuning performance comparison.** (A) Test F1 score distributions comparing base and fine-tuned models across all participants. (B) Absolute F1 improvement distribution showing consistent gains. (C) Relative improvement as percentage of base performance. (D) Improvement relative to room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

243 4.4 Fine-Tuning Outperforms Training From Scratch

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

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

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

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	$\Delta F1$	Rel. Imp. (%)	Precision	Recall	$\Delta Prec$	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

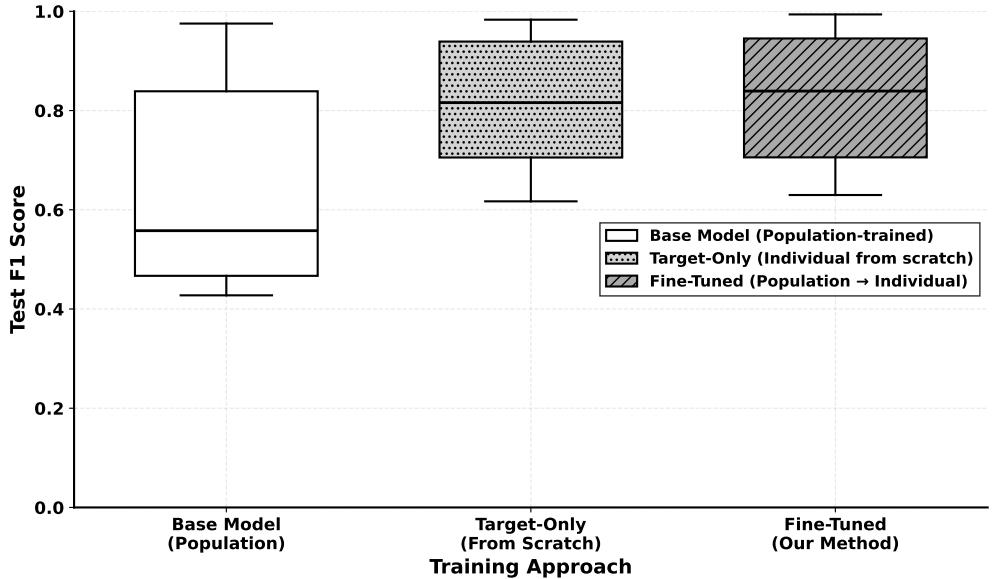


Figure 3: **Comprehensive performance comparison: three training approaches.** Test F1 scores for base model, target-only model, and fine-tuned model across all individuals, highlighting the superiority of fine-tuning. Even with 100% of individual data available to both approaches, fine-tuned models (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.

258 4.5 Transfer Learning Excels in Low-Data Regimes

259 The critical advantage of transfer learning emerges when individual data is limited. We systematically
260 evaluated both approaches across data-scarce regimes by training models with 1%, 5%, 10%, 25%, 50%,

261 and 100% of each participant’s data (Figure 4). The results reveal a dramatic divergence in data efficiency
262 between methods.

263 At the most extreme data scarcity (1% of target data, approximately 1.3 hours per participant), the
264 superiority of transfer learning is most pronounced. Target-only models trained from scratch collapsed to
265 median F1 of 0.535, barely exceeding chance performance and exhibiting high variance across participants.
266 In stark contrast, fine-tuned models maintained robust performance with median F1 of 0.627, demonstrating
267 an absolute improvement of 0.092 F1 points over target-only training. This improvement at 1% data is
268 nearly twice the magnitude of the mean improvement from base to fine-tuned models at full data (0.130),
269 underscoring the critical value of population-pretrained representations when individual labels are extremely
270 scarce. Remarkably, fine-tuning with just 1% of target data retains 74.2% of the performance achieved with
271 100% of target data, while target-only training at 1% achieves only 65.5% of its full-data performance—a
272 8.7 percentage point difference in retention. At this extreme, fine-tuned models with 1% data approach
273 the performance of base population models (median 0.627 vs. mean 0.647), demonstrating that minimal
274 individual data is sufficient to match or exceed generic population models.

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

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

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

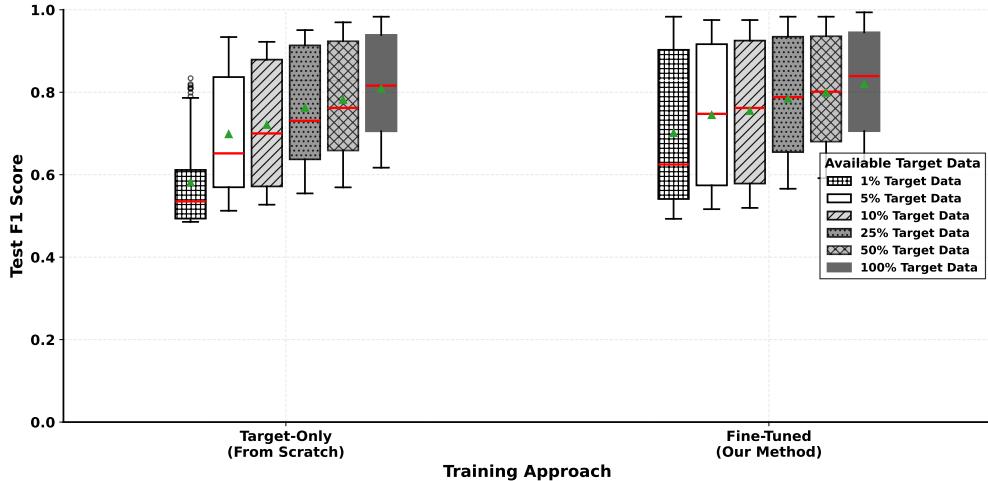


Figure 4: **Low-data regime performance comparison.** Test F1 scores for target-only and fine-tuned models across varying fractions of target data (1%, 5%, 10%, 25%, 50%, 100%), demonstrating the advantage of fine-tuning in data-scarce scenarios. The most extreme divergence occurs at 1% data, where fine-tuning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

294 4.6 Learned Feature Representations Balance Smoking Detection with Individ- 295 ual Variability

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

300 Feature space analysis demonstrates that fine-tuned models learn representations that simultaneously
301 capture shared smoking patterns and individual-specific variability. When colored by smoking label (Figure
302 5A), features exhibit clear separation between smoking and non-smoking windows, demonstrating that the
303 model successfully learned generalizable smoking behavior representations. This separation validates that
304 population pretraining captures meaningful gesture patterns common across individuals—the characteristic
305 hand-to-mouth motions, puffing dynamics, and temporal sequences that define smoking events.

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

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

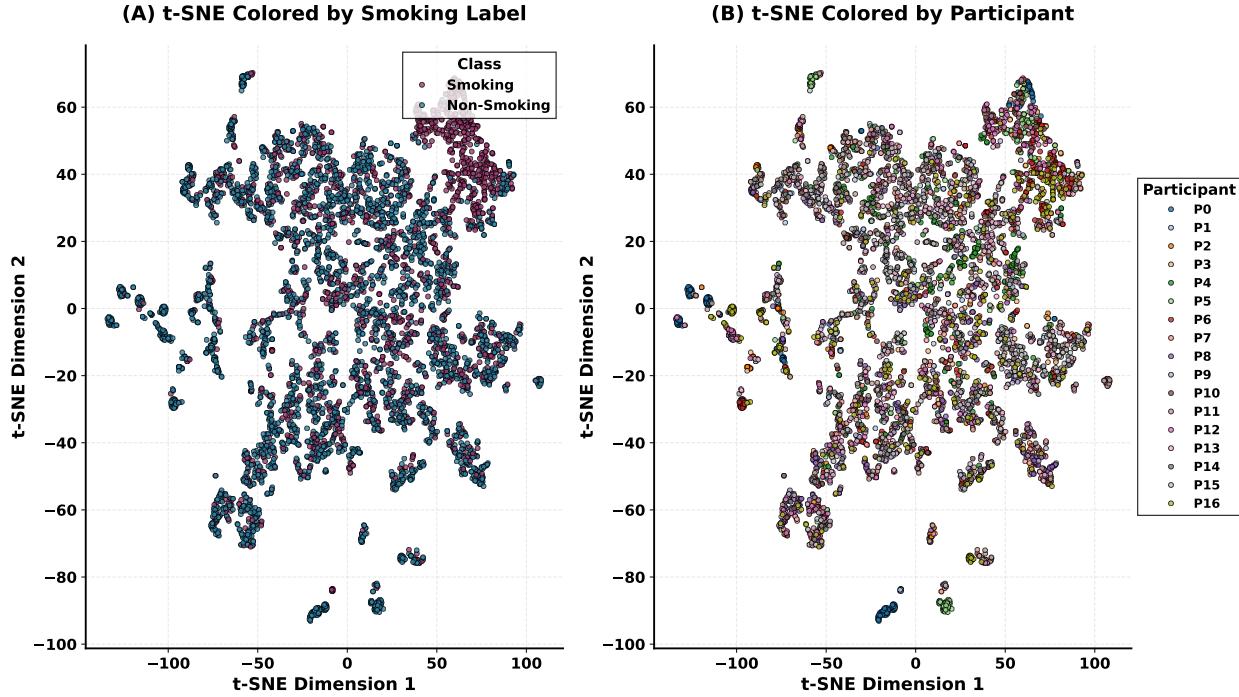


Figure 5: **Feature space visualization reveals dual structure of learned representations.** t-SNE projection of features extracted from a representative fine-tuned model applied to all participants’ test sets. (A) Features colored by smoking label show clear separation between smoking and non-smoking 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.

320 5 Discussion

321 We demonstrate that transfer learning resolves the fundamental tension between personalization accuracy
 322 and data collection burden in wearable behavioral sensing. By fine-tuning population-pretrained models with
 323 minimal individual data, we achieve robust personalized smoking detection with as little as 1.3 hours (1% of
 324 target data) of user data, where fine-tuning maintains median F1 of 0.627 while target-only training collapses
 325 to 0.535. This approach delivers mean F1 scores of 0.776 across 17 participants, substantially outperforming
 326 both population-only models (0.647) and data-matched individual models in low-data regimes.

327 5.1 Implications for Deployment

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

333 Most remarkably, our results demonstrate that even at the extreme of 1% individual data (approximately
 334 1.3 hours of wear time), fine-tuning achieves median F1 of 0.627—representing 74.2% of the performance
 335 obtained with full individual data collection. This represents a practical breakthrough: users can achieve
 336 highly functional personalized detection after wearing the device for barely more than one hour, rather

337 than 2 weeks. The absolute improvement of 0.092 F1 points over target-only training at this extreme
338 data scarcity underscores the critical value of population pretraining. For rapid deployment scenarios, pilot
339 studies, or populations where sustained data collection is challenging, the 1% regime provides a viable path
340 to personalization that was previously infeasible. At 5% data (\sim 6.4 hours), performance increases further,
341 providing deployment flexibility based on the accuracy-burden tradeoff appropriate for each use case.

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

347 5.2 Sources of Individual Variability

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

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

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

365 5.3 Comparison with Alternative Approaches

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

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

376 5.4 Limitations and Future Directions

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

382 Our models were trained and evaluated on wrist-worn accelerometer and gyroscope data from a specific
383 device. Generalization across different wearable form factors (arm bands, chest straps, rings) and sensor

384 modalities remains to be established. Cross-device transfer learning could extend our approach to heterogeneous
385 wearable ecosystems.

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

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

395 **5.5 Broader Impact**

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

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

406 **References**

- 407 [1] Volkan Y. Senyurek, Masudul H. Imtiaz, Prajakta Belsare, Stephen Tiffany, and Edward Sazonov.
408 A CNN-LSTM neural network for recognition of puffing in smoking episodes using wearable sensors.
409 *Biomedical Engineering Letters*, 10(2):195–203, January 2020.
- 410 [2] Casey A Cole, Dien Anshari, Victoria Lambert, James F Thrasher, and Homayoun Valafar. Detecting
411 Smoking Events Using Accelerometer Data Collected Via Smartwatch Technology: Validation Study.
412 *JMIR mHealth and uHealth*, 5(12):e189, December 2017.
- 413 [3] Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa.
414 Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. *Sensors*,
415 22(4):1476, January 2022.
- 416 [4] Sourish Gunesh Dhekane and Thomas Ploetz. Transfer Learning in Human Activity Recognition: A
417 Survey, January 2024.
- 418 [5] Personalized Human Activity Recognition Based on Integrated Wearable Sensor and Transfer Learning.
419 <https://www.mdpi.com/1424-8220/21/3/885>.
- 420 [6] Ali Akbari and Roozbeh Jafari. Personalizing Activity Recognition Models Through Quantifying Dif-
421 ferent Types of Uncertainty Using Wearable Sensors. *IEEE Transactions on Biomedical Engineering*,
422 67(9):2530–2541, September 2020.
- 423 [7] Megha Thukral, Harish Haresamudram, and Thomas Ploetz. Cross-Domain HAR: Few Shot Transfer
424 Learning for Human Activity Recognition, October 2023.
- 425 [8] Jianfei Yang, Yuecong Xu, Haozhi Cao, Han Zou, and Lihua Xie. Deep learning and transfer learning for
426 device-free human activity recognition: A survey. *Journal of Automation and Intelligence*, 1(1):100007,
427 December 2022.