Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SmartADAPT-Net: Decision-Level Adaptation for Free-Living Smartwatch-Based Activities of Daily Living Recognition

Mustafa Elhadi Ahmed

mahmed@qmu.ac.uk

Queen Margaret University https://orcid.org/0009-0009-3727-7330

Hongnian Yu

Edinburgh Napier University

Michael Vassallo

Royal Bournemouth Hospital

Pelagia Koufaki

Queen Margaret University

Research Article

Keywords: Smartwatch, ADL recognition, Wearable sensing, Adaptive recognition, Personalisation, Edge computing, CNN-BiLSTM

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10870927/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

SmartADAPT-Net: Decision-Level Adaptation for Free-Living Smartwatch-Based Activities of Daily Living Recognition

Mustafa Elhadi Ahmed[1,4*], Hongnian Yu[2], Michael Vassallo[3], Pelagia Koufaki[1]

[1*] School of Health Sciences, Queen Margaret University, Edinburgh, EH21 6UU, UK .
[2] School of Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, EH10 5DT, UK .
[3] Geriatric Medicine Department, Royal Bournemouth Hospital, Bournemouth, BH7 7DW, UK .
[4] School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK .

*Corresponding author(s). E-mail(s): mahmed@qmu.ac.uk;
Contributing authors: h.yu@napier.ac.uk; michael.vassallo@uhd.nhs.uk; PKoufaki@qmu.ac.uk;

1 Introduction

Activities of daily living (ADL) are commonly used to characterise an individual's level of functional independence [1]. Decline in ADL performance is clinically important because it is associated with adverse outcomes, including hospital readmission, mortality, and discharge to post-acute care [2, 3]. Recovery of ADL function can indicate improvement in functional status and inform rehabilitation and care planning [1, 4]. Conventional ADL assessment, however, is usually performed at discrete clinical or follow-up time points using standardised scales, questionnaires, or reported observations [1, 5]. Such assessments may be influenced by respondent perspective and recall [6, 7]. Repeated objective measurement in everyday settings could therefore complement conventional assessment by providing longitudinal information on functional behaviour [1, 8]. Consumer smartwatches provide a practical sensing platform for this

purpose because they combine routine wear with inertial sensing and onboard computation [9]. Smartwatch-based models have achieved high ADL-recognition accuracy under supervised and controlled conditions [10–12]. Performance established under structured conditions, however, does not ensure equivalent performance in unseen users or free-living environments.

Generalisation from structured development data to free-living behaviour remains a central problem in wearable human activity recognition (HAR). Controlled data-collection protocols typically restrict the activities performed, their temporal organisation, and the conditions under which they are executed [9, 13]. Free-living behaviour is less constrained and includes variability in activity duration, movement execution, transitions, environmental context, and co-occurring movements [13, 14]. Recent reviews have identified the transition from controlled benchmarks to unconstrained deployment as an important limitation of current HAR systems [15, 16]. Generalisation is further complicated by inter-subject variability, because the same activity can produce different inertial patterns across individuals [17]. These sources of variation can create a substantial mismatch between the data used for model development and those encountered during deployment.

The consequences of this mismatch have been demonstrated empirically. Stojchevska et al. reported a substantial reduction in balanced accuracy when laboratory-trained wrist-sensor models were evaluated using uncontrolled real-world recordings [18]. Burchard et al. similarly reported limited generalisation of a pretrained wrist-worn handwashing detector to previously unseen participants during prolonged real-world deployment [19]. Wrist placement also imposes a representational limitation. Wrist-worn inertial sensors capture upper-extremity movement but provide limited information about fine hand actions, object interaction, and surrounding context [20, 21]. Different ADLs can therefore produce overlapping wrist-motion patterns [22]. Free-living recognition must consequently address both distributional variation across users and environments and ambiguity in the information available from the sensing location.

Several learning strategies have been developed to improve generalisation beyond the original training distribution. These include transfer learning, domain adaptation, few-shot learning, test-time adaptation, and participant-specific personalisation [23–26]. These approaches differ principally in the amount and type of target-domain information they require. Few-shot methods adapt to new subjects or activities using a small labelled support set [27]. Personalised wearable recognition has similarly used limited target-user data for subject-specific adaptation or similarity-based classification [28–30]. Unsupervised and source-free approaches reduce the need for manually labelled target data but still rely on information obtained from the target distribution, for example through pseudo-labels, prediction consistency, updated normalisation, or adapted target representations [24, 26, 31, 32]. These approaches provide important mechanisms for adaptation once some form of target-domain information has become available.

Wearable deployment also differs from conventional offline adaptation because the information available to the system changes over time. At the start of deployment, there may be no participant-specific history. As use continues, repeated observations,

user feedback, and other forms of target-specific evidence may gradually become available. Treating deployment as a single adaptation event can therefore obscure an important temporal distinction between mechanisms that must operate immediately and those that can exploit evidence accumulated later. This distinction is particularly relevant to longitudinal wearable systems, where the amount and reliability of participant-specific information can evolve throughout deployment.

This temporal structure creates two different learning problems: how the population decision should respond when an unseen user is first encountered, and how reliable participant-specific information should subsequently be incorporated as it accumulates. Existing adaptation approaches largely address the latter problem once some form of target-domain information is available. For longitudinal wearable deployment, one complementary strategy is to preserve the learned population representation and restrict deployment-time changes to lightweight decision-level mechanisms. This avoids repeated gradient-based optimisation of the deployed network while allowing participant-specific information to be introduced progressively.

This study introduces SmartADAPT-Net, a smartwatch HAR framework designed around these two stages of deployment response. Complexity-Aware Modulation (CAM) provides observation-conditioned population modulation by using the motion complexity of each incoming inertial segment to regulate a fixed population-level correction. CAM is available from the first deployment recording and does not require participant-specific history. Confirmation-Driven Personalisation (CDP) subsequently incorporates labelled evidence obtained when participants confirm or correct the system prediction, allowing participant- and class-specific representations to accumulate progressively. CDP is therefore a human-in-the-loop personalisation mechanism rather than an unsupervised adaptation approach. The two mechanisms are combined so that population-level support is retained while participant-specific evidence is progressively incorporated.

The study addresses two questions: (1) what change in free-living recognition performance is observed after complexity-aware population modulation, before participant-specific evidence is available; and (2) what additional change is observed as participant-confirmed evidence accumulates? Activity-specific residual errors are examined as a secondary analysis.

2 Methods

2.1 Study design

This study used a two-phase design to develop and evaluate smartwatch-based ADL recognition across structured model development and participant-initiated free-living deployment. Phase 1 was used to develop and evaluate a participant-independent population model using labelled smartwatch recordings collected in a simulated-living environment. Phase 2 evaluated the resulting framework in a separate cohort during 20 days of use in participants' everyday environments.

SmartADAPT-Net extends the population model through two decision-level mechanisms. Complexity-Aware Modulation (CAM) applies an observation-conditioned correction to the population prediction using the motion complexity of the current

inertial window. Confirmation-Driven Personalisation (CDP) incorporates participant-confirmed observations into participant- and class-specific prototype memory as deployment progresses. The neural encoder, population classifier, CAM correction head, normalisation parameters, and complexity calibration remained fixed throughout Phase 2. Deployment-time updates were restricted to participant-specific prototype vectors and their confirmation counts.

The Phase 1 and Phase 2 cohorts were disjoint, and no participant contributed data to both phases. The cohorts were not demographically matched. Differences between the two phases therefore represent a combined change in participant population, recording environment, activity execution, and evaluation setting rather than an isolated estimate of environmental domain shift.

Figure 1 summarises the framework. The population pathway comprises the frozen CNN–BiLSTM classifier and CAM. The personalisation pathway compares the current representation with participant-specific class prototypes. The resulting probability distributions are combined through evidence-adaptive fusion. Participant feedback is incorporated only after the prediction for the current recording has been stored and can therefore influence subsequent recordings only.

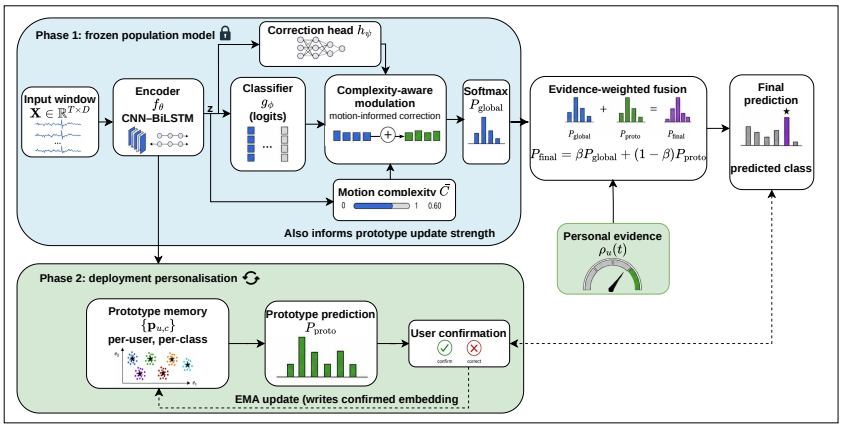


Fig. 1 Overview of SmartADAPT-Net. The population recognition model and Complexity-Aware Modulation pathway are fixed before deployment. Participant-confirmed observations update class-specific prototype memory in the frozen embedding space and contribute progressively through evidence-adaptive fusion. Feedback from the current recording affects subsequent predictions only.

2.2 Ethical approval and participants

Ethical approval was granted by the Research Ethics Panel of the authors' institution under reference REP0278 for the study "Wearable devices for supporting frail individuals to live independently" on 19 December 2022. All study procedures were conducted in accordance with the relevant institutional guidelines and regulations. All participants provided written informed consent before participation.

Phase 1 participants were adults able to perform the six target ADLs independently and to wear and use the study smartwatch. Phase 2 participants met the same functional requirements and were also required to operate the smartwatch application during deployment. Individuals were excluded where a physical, cognitive, or communication limitation prevented safe completion of the study procedures or provision of informed consent.

Participant demographic characteristics are reported in the Results section.

2.3 Phase 1: simulated-living data collection

Phase 1 data were collected in a domestic-style simulated-living environment. The setting reproduced common household contexts while allowing activities to be performed under structured, researcher-observed conditions.

Thirty participants completed six target ADLs: Brushing Teeth, Mopping/Hoovering, Shelving Items, Walking, Washing Dishes, and Washing Face. Each activity was performed for approximately 2 minutes, and activity order was randomised between participants.

Motion data were recorded using an Apple Watch Series 8 through the CoreMotion framework. Tri-axial user acceleration and tri-axial angular velocity were acquired at an effective sampling frequency of approximately 50 Hz. A custom smartwatch application was used to initiate and terminate recordings. Phase 1 labels were assigned according to the researcher-observed activity performed under the structured protocol.

Phase 1 data were used to train and evaluate the participant-independent population model, estimate inertial-channel normalisation statistics, estimate motion-complexity descriptor statistics, and calibrate the population complexity distribution used by CAM.

2.4 Input construction and recording-level aggregation

The recognition model used tri-axial user acceleration and tri-axial gyroscope signals, giving six inertial channels per model input. Signals were sampled at approximately 50 Hz. The model operated on contiguous non-overlapping 8-second windows containing 400 samples:

$$\mathbf{X}_j \in \mathbb{R}^{400\times 6}, \tag{1}$$

where j indexes a complete analysis window.

Phase 1 activity recordings were divided into consecutive non-overlapping 8-second windows. Only complete windows were used for model development and evaluation; incomplete trailing segments were discarded.

During Phase 2, each participant-initiated recording began with a 3-second stabilisation period, which was excluded from recognition. The subsequent activity recording was divided into consecutive non-overlapping 8-second windows. At least one complete 8-second window was required for classification. Any residual segment shorter than 8 seconds did not contribute to the recording-level decision.

Each complete window was processed independently through the recognition framework. For a recording containing M complete windows, the probability distribution for decision stage q was aggregated by arithmetic averaging:

$$P_{\text{rec}}^{(q)}(y = c) = \frac{1}{M} \sum_{j=1}^{M} P_j^{(q)}(y = c), \tag{2}$$

and the corresponding recording-level prediction was

$$\hat{y}_{\text{rec}}^{(q)} = \arg\max_c P_{\text{rec}}^{(q)}(y = c). \tag{3}$$

This procedure constitutes probability averaging, or soft voting, rather than majority voting over hard window-level class labels. Consequently, each participant-initiated recording generated one recording-level prediction irrespective of the number of complete 8-second windows available.

The six inertial channels were z-score normalised using statistics estimated exclusively from the corresponding Phase 1 training data. During participant-independent cross-validation, normalisation statistics were estimated from training participants only and applied unchanged to held-out participants. The same leakage-prevention rule was applied to all motion-complexity calibration parameters.

2.5 Population recognition model

The population model used a CNN–BiLSTM architecture to combine local temporal feature extraction with recurrent modelling. Three one-dimensional convolutional blocks used 64, 96, and 128 filters, with kernel sizes 7, 5, and 3 and strides 1, 2, and 2, respectively. Each convolution was followed by batch normalisation, rectified linear activation, and dropout with probability 0.20.

The convolutional representation was passed to a bidirectional LSTM with 64 units in each direction. A dense layer produced a 128-dimensional embedding, followed by dropout with probability 0.30. For an input window $\mathbf{X}_j$, the encoder f_θ produced an embedding $\mathbf{z}_j$, and the population classifier g_ϕ produced Base-stage logits $\mathbf{s}_j$:

$$\mathbf{z}_j = f_\theta(\mathbf{X}_j), \qquad \mathbf{s}_j = g_\phi(\mathbf{z}_j). \tag{4}$$

Base-stage probabilities were

$$P_{\text{base},j} = \text{softmax}(\mathbf{s}_j). \tag{5}$$

The 128-dimensional embedding was retained for both CAM and the participant-specific prototype pathway.

2.6 Phase 1 training and participant-independent evaluation

Phase 1 evaluation used participant-independent 5-fold `GroupKFold` cross-validation, with participant identity as the grouping variable. All windows from an individual participant were retained within the same fold, preventing participant overlap between training and validation data.

Within each fold, inertial-channel normalisation statistics and motion-complexity calibration parameters were estimated exclusively from training participants. The encoder and Base classifier were first trained jointly using sparse categorical cross-entropy applied to the Base logits. Training used the Adam optimiser with an initial learning rate of 10^{-3}, batch size 64, and a maximum of 120 epochs. Early stopping monitored validation accuracy with a patience of 15 epochs. The learning rate was reduced by a factor of 0.5 after seven epochs without improvement, to a minimum of 10^{-6}.

After Base training, the encoder and population classifier were frozen. The CAM correction head was then trained while leaving the population representation unchanged. The correction head comprised a 64-unit rectified-linear dense layer with dropout probability 0.20 followed by a six-dimensional logit correction layer. It was optimised using cross-entropy on the CAM-adjusted logits, with an initial learning rate of 5×10^{-4} and a maximum of 80 epochs using the same early-stopping procedure.

After participant-independent evaluation, a separate deployment model was trained using the Phase 1 data with participant-level validation used during optimisation. The deployment weights were therefore distinct from the fold-specific models used to generate the participant-independent Phase 1 performance estimates. The resulting network and all population-level preprocessing and complexity-calibration parameters were fixed before Phase 2. No neural-network parameter was retrained using Phase 2 data.

2.7 Complexity-Aware Modulation

CAM provides observation-conditioned population modulation without requiring participant-specific deployment history. The term *complexity* is used operationally to describe the dynamic irregularity of an incoming inertial window relative to the Phase 1 reference distribution and does not denote an information-theoretic complexity measure.

For a generic tri-axial signal $\mathbf{v}$ of length T, three descriptors were calculated: mean absolute magnitude, dynamic range, and mean absolute second temporal difference. These were computed for both accelerometer and gyroscope channels and z-score normalised using Phase 1 training statistics.

The normalised accelerometer descriptors were averaged to obtain c_{acc}, and the corresponding gyroscope descriptors were averaged to obtain c_{gyro}. The raw complexity score was

$$C = w_{\mathrm{acc}} c_{\mathrm{acc}} + (1 - w_{\mathrm{acc}}) c_{\mathrm{gyro}}, \qquad w_{\mathrm{acc}} = 0.4. \tag{6}$$

The raw score was calibrated relative to the Phase 1 distribution:

$$\bar{C} = \mathrm{clip}\left(\frac{C - q_{15}}{q_{85} - q_{15}}, 0, 1 \right), \tag{7}$$

where q_{15} and q_{85} are the 15th and 85th percentiles of the Phase 1 training complexity distribution.

The calibrated score controlled a sigmoid gate:

$$g(\bar{C}) = \sigma\left[k(\bar{C} - \tau)\right], \tag{8}$$

with $\tau = 0.5$ and $k = 10$.

A correction head h_ψ operated on the frozen embedding and produced a six-dimensional residual logit vector:

$$\boldsymbol{\Delta}_j = h_\psi(\mathbf{z}_j). \tag{9}$$

The complexity score did not enter the correction head directly. Instead, it scaled the learned correction before addition to the Base logits:

$$\widetilde{\mathbf{s}}_j = \mathbf{s}_j + g(\bar{C}_j)\boldsymbol{\Delta}_j, \tag{10}$$

giving

$$P_{\text{global},j} = \text{softmax}\left(\widetilde{\mathbf{s}}_j\right). \tag{11}$$

The correction head, complexity calibration, and gate parameters were fixed before Phase 2. CAM was therefore available for the first deployment recording without participant identity, confirmation history, participant-specific calibration, or deployment-time gradient optimisation.

The Base-to-Global comparison evaluates the CAM pathway as an integrated mechanism comprising both the trained residual correction head and complexity-dependent gating. The present analysis does not separately estimate the independent contribution of the correction head and the complexity gate.

2.8 Confirmation-Driven Personalisation

CDP introduced participant-specific information while preserving the population embedding space, as illustrated in Figure 2. Neither the encoder, population classifier, nor CAM correction head was modified by participant feedback.

For participant u and class c, CDP maintained a prototype vector $\mathbf{p}_{u,c}$ and confirmation count $n_{u,c}$. Both were initialised to zero. Window embeddings were ℓ_2-normalised before prototype operations:

$$\bar{\mathbf{z}}_j = \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}. \tag{12}$$

A class prototype became active when

$$n_{u,c} \geq N_{\min}, \qquad N_{\min} = 3. \tag{13}$$

For each window, cosine similarity between the current normalised embedding and active participant-specific prototypes was converted into a probability distribution:

$$P_{\text{proto},j}(y = c \mid u, t) = \frac{A_{u,c}(t)\exp\left(\lambda \bar{\mathbf{z}}_j^\top \mathbf{p}_{u,c}\right)}{\sum_{c'} A_{u,c'}(t)\exp\left(\lambda \bar{\mathbf{z}}_j^\top \mathbf{p}_{u,c'}\right)}, \tag{14}$$

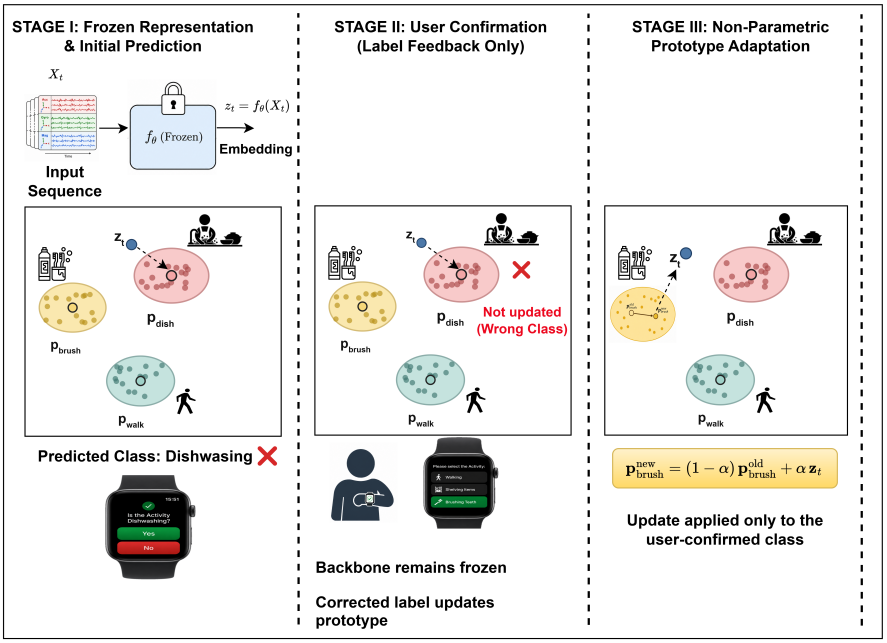


Fig. 2 Confirmation-Driven Personalisation (CDP). Participant-specific class prototypes are updated from confirmed or corrected recordings in the frozen embedding space and influence subsequent predictions.

where $A_{u,c}(t)$ indicates whether the prototype was active before recording t, and $\lambda = 10$. Inactive classes therefore contributed zero probability once at least one prototype was active. When no prototype was active, the prototype distribution was defined as uniform but excluded from the Final decision by the fusion mechanism.

Prototype memory was updated only after the recording-level prediction had been stored and the participant had confirmed or corrected the activity label. To form a single representation for a recording containing M complete windows, the normalised window embeddings were combined using their Final-stage probabilities for the participant-confirmed class:

$$\bar{\mathbf{z}}_{\text{rec}} = \frac{\sum_{j=1}^{M} P_{\text{final},j}\left(y = c_t^{\text{conf}}\right)\bar{\mathbf{z}}_j}{\sum_{j=1}^{M} P_{\text{final},j}\left(y = c_t^{\text{conf}}\right)}, \tag{15}$$

after which the resulting vector was renormalised to unit norm.

The recording-level calibrated complexity used for prototype updating was the arithmetic mean across complete windows:

$$\bar{C}_{\mathrm{rec}} = \frac{1}{M}\sum_{j=1}^{M}\bar{C}_j. \tag{16}$$

The prototype update rate was

$$\alpha_t = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})(1 - \bar{C}_{\mathrm{rec}}), \tag{17}$$

where

$$\alpha_{\min} = 0.05, \qquad \alpha_{\max} = 0.30. \tag{18}$$

For the participant-confirmed class c_t^{conf}, the prototype was updated using an exponential moving average:

$$\mathbf{p}^{*}_{u,c} = (1-\alpha_t)\mathbf{p}^{(t)}_{u,c} + \alpha_t \bar{\mathbf{z}}_{\mathrm{rec}}, \qquad c = c_t^{\mathrm{conf}}, \tag{19}$$

followed by unit normalisation:

$$\mathbf{p}^{(t+1)}_{u,c} = \frac{\mathbf{p}^{*}_{u,c}}{\|\mathbf{p}^{*}_{u,c}\|_2}. \tag{20}$$

The corresponding confirmation count was incremented by one. All other prototypes and counters remained unchanged. If the participant rejected the predicted activity and selected a different class, only the corrected-class prototype was updated; the rejected predicted-class prototype remained unchanged.

2.9 Evidence-adaptive fusion

The Final-stage prediction combined the CAM-adjusted population distribution with the participant-specific prototype distribution. Prototype influence increased progressively as participant-specific confirmation evidence accumulated.

For participant u, confirmation coverage at time t was

$$\rho_u(t) = \frac{1}{K}\sum_{c=1}^{K}\min\left(\frac{n_{u,c}(t)}{N_{\mathrm{ref}}}, 1\right), \tag{21}$$

where $K = 6$ and $N_{\mathrm{ref}} = 5$.

The nominal population weight was

$$\beta_u(t) = 1 - \gamma\rho_u(t), \tag{22}$$

with $\gamma = 0.5$. Participant-specific information could therefore contribute at most half of the Final decision weight.

To prevent an uninformative prototype distribution from influencing the system before any prototype was active, the effective population weight was

$$\beta_u^{\mathrm{eff}}(t) = \begin{cases} 1, & \text{if no participant-specific prototype is active,} \\ \beta_u(t), & \text{otherwise.} \end{cases} \tag{23}$$

The window-level Final distribution was then

$$P_{\text{final},j}(y = c \mid u, t) = \beta_u^{\text{eff}}(t) P_{\text{global},j}(y = c) + \left[1 - \beta_u^{\text{eff}}(t)\right] P_{\text{proto},j}(y = c \mid u, t). \quad (24)$$

These window-level distributions were subsequently aggregated by soft voting according to Eq. 2. At cold start, $\beta_u^{\text{eff}}(t) = 1$, so the Final prediction was identical to the CAM-adjusted population pathway. As confirmations accumulated, participant-specific prototype evidence contributed progressively while the population pathway retained at least half of the fusion weight.

2.10 Hyperparameter specification

All CAM, prototype, and fusion hyperparameters were fixed before Phase 2 evaluation. Phase 2 outcomes were not used to select or optimise the reported configuration.

The predefined parameters were $w_{\text{acc}} = 0.4$, $\tau = 0.5$, $k = 10$, $N_{\min} = 3$, $\lambda = 10$, $\alpha_{\min} = 0.05$, $\alpha_{\max} = 0.30$, $N_{\text{ref}} = 5$, and $\gamma = 0.5$.

These parameters define the implemented SmartADAPT-Net configuration evaluated in Phase 2. No claim is made that the individual values constitute globally optimal hyperparameters.

2.11 Phase 2: participant-initiated free-living deployment

The trained framework was converted to CoreML and integrated into a watchOS application deployed on an Apple Watch Series 8. A separate cohort of 17 participants used the system for 20 days during their normal daily routines.

Participants were instructed to record the same six ADLs considered during Phase 1 where feasible and were asked to perform each activity approximately twice per day to support class coverage. Activity timing, environmental context, object use, and execution strategy were not experimentally standardised.

Phase 2 evaluated participant-initiated classification in a free-living setting rather than autonomous recognition from a continuous background sensor stream. Participants initiated a recording through the smartwatch immediately before performing one of the target activities. Following the 3-second stabilisation period, the subsequent activity recording was analysed using complete 8-second windows as described above.

Recognition was closed-set among the six target ADLs. Background activity detection, automatic activity-boundary identification, out-of-vocabulary recognition, and rejection of activities outside the six trained classes were outside the scope of the Phase 2 evaluation.

For each completed recording, the system generated one recording-level prediction using soft voting across the complete 8-second windows. The predicted activity was then displayed to the participant. Selecting "Yes" confirmed the displayed class. Selecting "No" opened the six-class activity list, from which the participant selected the activity they had performed.

Each retained recording therefore contained the inertial data, model prediction, and a participant-confirmed or corrected deployment reference label. Because participants viewed the system prediction before confirming or correcting it, these labels

were treated as participant-confirmed deployment reference labels rather than independently observer-verified ground truth. The procedure is therefore susceptible to confirmation or automation bias of unknown magnitude. Because the same confirmations were subsequently used to update prototype memory, erroneous confirmations could also influence later participant-specific state.

Feedback associated with recording t was incorporated only after the prediction for recording t had been stored. It could therefore influence recording $t+1$ and subsequent recordings but could not alter the prediction from which the feedback had been obtained.

2.12 Chronological deployment-log analysis

Phase 2 recordings were analysed chronologically and separately for each participant. Prototype memory was initialised as empty at the beginning of each participant's deployment history and reconstructed sequentially. For every recording, all decision stages were calculated using only participant-specific evidence available before that recording. The participant-confirmed or corrected label from the current recording was incorporated only after its prediction had been stored.

No future-recording labels, Phase 2 neural-network optimisation, or backbone updates were permitted. This preserved the information state available at each point in deployment and prevented future-label leakage.

Four decision stages were retained:

Table 1 Decision stages used in the chronological deployment-log analysis.

Stage	**Available evidence**	**Participant-specific**
Base	Frozen CNN–BiLSTM population prediction before CAM	No
Global	Population pathway after the complete CAM correction	No
Prototype	Similarity to active participant- and class-specific prototypes	Yes
Final	Evidence-adaptive fusion of Global and Prototype probabilities	Yes

Prototype-stage performance was evaluated only for recordings in which at least one participant-specific prototype had reached the activation threshold. Prototype-stage metrics therefore describe a selected prototype-active subset and are not directly comparable with Base, Global, or Final metrics calculated over the complete Phase 2 recording set. Base, Global, and Final predictions were retained for all eligible recordings, including those occurring during prototype cold start.

The chronological stage analysis was descriptive. Base-to-Global differences represent the observed change after the complete CAM pathway and do not isolate

the independent effects of the correction head and complexity gate. Global-to-Final differences represent the additional stage-wise change observed after participant-specific prototype evidence became available. Neither comparison was interpreted as an independent causal effect.

2.13 Prediction-transition analysis

Each Phase 2 recording was additionally classified according to the relationship between its Base and Final predictions:

- *Stable correct*: both Base and Final matched the deployment reference label;
- *Recovered*: Base was incorrect and Final matched the deployment reference label;
- *Degraded*: Base matched the deployment reference label and Final was incorrect; and
- *Persistent failure*: both Base and Final were incorrect.

Recovery efficiency was defined descriptively as

$$\text{Recovery efficiency} = \frac{N_{\text{recovered}}}{N_{\text{Base errors}}}, \tag{25}$$

and the degradation rate as

$$\text{Degradation rate} = \frac{N_{\text{degraded}}}{N_{\text{Base correct}}}. \tag{26}$$

Net correction was defined as

$$N_{\text{net}} = N_{\text{recovered}} - N_{\text{degraded}}. \tag{27}$$

These quantities describe transitions between Base and Final classification outcomes relative to the same participant-confirmed deployment reference label. They do not imply independence between the two decision stages and were not interpreted as causal effects of individual adaptive components.

2.14 Outcome measures and statistical analysis

The primary Phase 2 outcome was Final-stage recording-level recognition accuracy during participant-initiated free-living deployment. Macro-averaged F1 score and balanced accuracy were additionally reported to account for class-specific performance and unequal class support. Precision, recall, F1 score, and support were reported for individual ADLs.

Performance was characterised for the Base, Global, Prototype, and Final stages. Prototype-stage metrics were calculated only for prototype-active recordings and were reported with their corresponding sample size. They were treated as diagnostic metrics for the prototype-active subset rather than as directly comparable estimates against full-recording stage metrics.

Stage-wise differences were examined descriptively at the overall, activity, and participant levels. The Base-to-Global difference characterised the change present after the complete CAM pathway and before participant-specific prototype evidence contributed to the Final decision. The subsequent Global-to-Final difference characterised

the additional stage-wise change after personal evidence became available. These differences were not interpreted as independent causal effects.

Uncertainty in the principal Phase 2 estimates was quantified using a participant-clustered non-parametric bootstrap with $B = 10{,}000$ resamples. In each bootstrap iteration, 17 participants were sampled with replacement, and all recordings belonging to each sampled participant were retained. This preserved the repeated-measure structure within participants. The 2.5th and 97.5th percentiles of the bootstrap distribution defined the 95% confidence interval.

Participant-level accuracy was also calculated separately for each participant to assess the consistency of stage-wise performance patterns across the deployment cohort.

2.15 On-device computational assessment

Computational performance was evaluated on an Apple Watch Series 8 using the same CoreML and watchOS implementation used during Phase 2. The assessment considered model inference latency, recording-processing latency, peak process memory, participant-specific prototype-memory requirements, processor behaviour, and battery consumption.

Model inference latency was measured after model warm-up over 500 repeated executions of the 8-second model input. Complete recording-level processing latency was evaluated over 100 repetitions of the deployed processing pipeline, including preprocessing, window-level population inference, CAM, prototype comparison where available, fusion, and recording-level probability aggregation.

Runtime memory and processor behaviour were assessed using Xcode Instruments. Prototype-memory overhead was calculated for six 128-dimensional single-precision participant-specific class prototypes together with their associated state variables.

Battery consumption was evaluated under controlled one-hour conditions comparing sensing with recognition against sensing without recognition. Each condition was repeated five times using the same watch and watchOS version and an identical recording schedule. Runs were conducted at an ambient temperature of $22 \pm 1\,^{\circ}\mathrm{C}$. Airplane Mode was enabled, wireless interfaces and the always-on display were disabled, and the screen remained off during the benchmark. Battery consumption was expressed as the incremental percentage-point depletion associated with the recognition pipeline.

These measurements were intended to characterise controlled computational overhead and technical feasibility on the target wearable device. They were not interpreted as estimates of unrestricted all-day battery life or evidence of clinical effectiveness.

3 Results

3.1 Participant and dataset characteristics

Participant characteristics for both study phases are summarised in Table 2. The Phase 1 cohort included 30 participants and the Phase 2 cohort included 17 participants. The Phase 1 cohort was older on average and included a higher proportion of

female participants than the Phase 2 cohort. Across the 20-day Phase 2 deployment, 2523 participant-initiated ADL recordings were available for analysis.

Table 2 Participant characteristics in the simulated-living and free-living phases. Values are mean ± SD or number (percentage).

Characteristic	Phase 1	Phase 2
Participants, n	30	17
Age, years	56.6 ± 17.3	46.1 ± 15.5
Female, n (%)	16 (53.3%)	5 (29.4%)
Male, n (%)	14 (46.7%)	12 (70.6%)
Height, cm	172.4 ± 6.6	–
Body mass, kg	71.5 ± 7.6	–
Body mass index, kg/m^2	24.0 ± 2.5	–
Deployment recordings, n	–	2523

Height, body mass, and body mass index were available for Phase 1 participants only.

3.2 Population-model performance in the simulated living environment

The participant-independent population model was first evaluated using the Phase 1 simulated-living recordings. The frozen CNN–BiLSTM Base model achieved an overall window-level accuracy of 0.90 and an overall F1-score of 0.89 (Table 3). Performance was highest for Mopping/Hoovering, Shelving Items, and Walking, with class-wise accuracies between 0.95 and 0.96. Washing Dishes achieved an accuracy of 0.93 and an F1-score of 0.90. Brushing Teeth achieved an accuracy of 0.81 and an F1-score of 0.87, while Washing Face showed the lowest performance, with an accuracy of 0.74 and an F1-score of 0.66.

After applying the complete complexity-aware population pathway, overall accuracy increased from 0.90 to 0.92 and overall F1-score from 0.89 to 0.90. The direction and magnitude of the stage-wise difference varied across activities. Washing Dishes increased from 0.93 to 0.95 in accuracy, Brushing Teeth from 0.81 to 0.83, Shelving Items from 0.95 to 0.97, Walking from 0.95 to 0.96, and Washing Face from 0.74 to 0.75. Mopping/Hoovering was the only activity for which accuracy decreased, from 0.96 to 0.94.

Figure 3 shows the corresponding class-wise confusion patterns.

3.3 Free-living deployment performance

Across the 20-day Phase 2 deployment, 17 participants contributed 2523 participant-initiated ADL recordings. Each recording generated one recording-level prediction through probability averaging across all complete 8-second analysis windows. Of the 2523 recordings, 2490 occurred after at least one participant-specific prototype had become active, while 33 occurred during prototype cold start.

Table 3 Participant-independent Phase 1 performance of the frozen CNN–BiLSTM Base model and the complete complexity-aware population pathway.

ADL	Base			Complexity-aware pathway		
2-7	Accuracy	Precision	F1-score	Accuracy	Precision	F1-score
Brushing Teeth	0.81	0.93	0.87	0.83	0.93	0.88
Washing Dishes	0.93	0.88	0.90	0.95	0.90	0.92
Mopping/Hoovering	0.96	0.95	0.96	0.94	0.96	0.95
Shelving Items	0.95	0.95	0.95	0.97	0.96	0.96
Walking	0.95	0.98	0.96	0.96	0.98	0.97
Washing Face	0.74	0.61	0.66	0.75	0.63	0.68
Overall	0.90	–	0.89	0.92	–	0.90

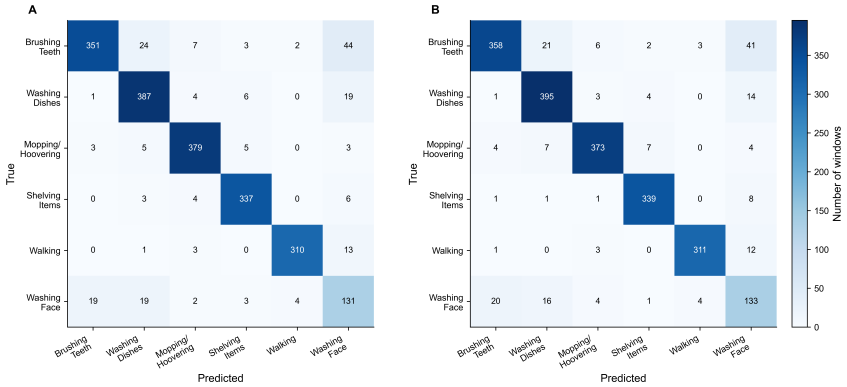


Fig. 3 Confusion matrices for the frozen CNN–BiLSTM Base model and the complete complexity-aware population pathway under participant-independent Phase 1 evaluation.

The frozen Base stage achieved an overall recording-level accuracy of 0.621 during Phase 2, compared with the window-level Phase 1 Base accuracy of 0.900. This difference reflects the combined change in participant cohort, recording environment, activity execution, and evaluation unit and should not be interpreted as an isolated estimate of environmental domain shift.

The Global/CAM stage achieved an overall accuracy of 0.784. The Final fused stage achieved an accuracy of 0.890 (participant-clustered 95% CI, 0.869–0.908), a macro-F1 of 0.892, and a balanced accuracy of 0.892. These Phase 2 performance estimates were calculated relative to participant-confirmed or corrected deployment reference labels and were not independently observer-verified.

Table 4 shows the stage-wise results for each ADL. Final-stage accuracy was highest for Walking and Mopping/Hoovering, both reaching 0.940. Washing Dishes reached 0.915, followed by Shelving Items at 0.890 and Brushing Teeth at 0.880. Washing Face remained the lowest-performing activity at 0.790, despite showing the largest absolute Base-to-Final difference.

Table 4 Recording-level Phase 2 performance and stage-wise accuracy differences by ADL. Base-to-Global is Global/CAM minus Base, while Global-to-Final is Final minus Global/CAM.

ADL	n	Base	Global/CAM	Final	Base→Global	Global→Final
Brushing Teeth	408	0.600	0.770	0.880	0.169	0.110
Mopping/Hoovering	384	0.781	0.870	0.940	0.089	0.070
Shelving Items	481	0.620	0.780	0.890	0.160	0.110
Walking	417	0.779	0.871	0.940	0.091	0.070
Washing Dishes	376	0.649	0.819	0.915	0.170	0.096
Washing Face	457	0.339	0.619	0.790	0.280	0.171
Overall	2523	0.621	0.784	0.890	0.163	0.106

The all-recording confusion matrices (Figure 4) show the progression from Base to Global/CAM and Final. The Global stage showed higher class-wise accuracy than Base for all six Phase 2 ADLs, and the Final stage showed the highest accuracy for every ADL. Washing Face Base-stage errors were concentrated primarily toward Mopping/Hoovering and, to a lesser extent, Walking. Although this class showed the largest absolute improvement, residual confusion remained at the Final stage.

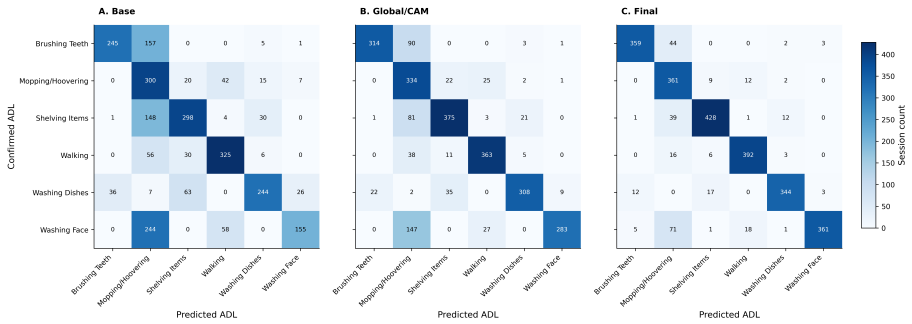


Fig. 4 Recording-level confusion matrices for the Base, Global/CAM, and Final stages using the same 2523 Phase 2 recordings. The Prototype stage is not shown because it was defined only for the 2490 prototype-active recordings.

3.4 Recovery before and after participant-specific evidence

The chronological deployment analysis separated the decisions available before participant-specific prototype evidence from those made after such evidence became available. The Base stage achieved an overall accuracy of 0.621, while Global/CAM reached 0.784, corresponding to an observed Base-to-Global difference of 0.163.

Because the Global stage contains both the trained residual correction head and its complexity-dependent gate, this difference represents the complete CAM pathway and does not isolate the independent contribution of complexity gating.

The Final stage reached 0.890, corresponding to an additional Global-to-Final difference of 0.106. Approximately 61% of the observed Base-to-Final accuracy difference was therefore already present at the Global stage.

This stage-wise decomposition describes where the observed accuracy differences appeared within the framework and should not be interpreted as a causal partition of the effects of individual adaptive components.

The Prototype branch alone achieved an accuracy of 0.850 on the 2490 prototype-active recordings. Prototype-stage performance is conditional on prior prototype activation and therefore characterises a selected, predominantly post-cold-start subset. For descriptive comparison, Base accuracy restricted to the same 2490 recordings was 0.621.

For the remaining 33 cold-start recordings, no participant-specific prototype was active. The effective fusion weight was therefore equal to one, and the Final probability distribution was identical to the Global distribution by construction.

The relative size of the stage-wise differences varied across ADLs (Figure 5). Washing Face showed the largest Base-to-Global difference at 0.280, although this occurred from the lowest Base-stage accuracy of 0.339. Washing Dishes, Brushing Teeth, and Shelving Items showed differences of 0.170, 0.169, and 0.160, respectively. Walking and Mopping/Hoovering showed smaller differences of 0.091 and 0.089. The subsequent Global-to-Final difference was positive for all six activities and ranged from 0.070 for Walking and Mopping/Hoovering to 0.171 for Washing Face.

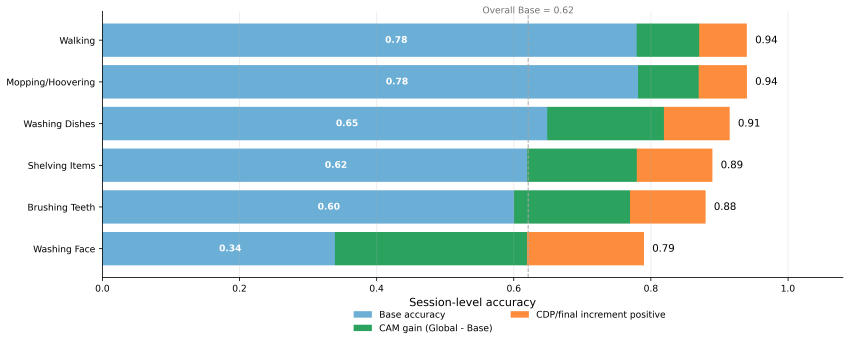


Fig. 5 Stage-wise decomposition of the observed Base-to-Final accuracy difference into the Base-to-Global difference after the complete CAM pathway and the subsequent Global-to-Final difference after participant-specific prototype evidence was permitted to contribute. The segments describe sequential decision-stage differences and are not interpreted as independent causal effects.

Table 5 summarises the integrated decision-stage performance. Macro-F1 increased from 0.635 at Base to 0.793 at Global and 0.892 at Final. Balanced accuracy increased from 0.628 to 0.788 and 0.892, respectively.

Participant-level mean accuracy increased from 0.622 ± 0.071 at Base to 0.781 ± 0.071 at Global and 0.887 ± 0.047 at Final. Positive overall Base-to-Final differences were observed for all 17 participants.

Table 5 Integrated decision-stage performance during Phase 2 free-living deployment. Confidence intervals were obtained using participant-clustered bootstrapping.

Analysis	Base CNN–BiLSTM	Global/CAM	Final SmartADAPT-Net
Recording accuracy, 95% CI	0.621 (0.586–0.652)	0.784 (0.750–0.813)	0.890 (0.869–0.908)
Macro-F1, 95% CI	0.635 (0.594–0.668)	0.793 (0.759–0.823)	0.892 (0.871–0.909)
Balanced accuracy, 95% CI	0.628 (0.592–0.660)	0.788 (0.755–0.817)	0.892 (0.872–0.909)
Participant accuracy, mean ± SD	0.622 ± 0.071	0.781 ± 0.071	0.887 ± 0.047
Participants with positive Base-to-Final difference	–	–	17/17
Base errors recovered	956 Base errors	–	769 recovered
Error-recovery proportion	–	–	0.804
Base-correct recordings becoming incorrect	1567 Base-correct recordings	–	91
Correct-to-incorrect transition proportion	–	–	0.058
Net change in correct decisions	–	–	+678 recordings

3.5 Error recovery and participant-level consistency

Of the 956 recordings misclassified by the Base stage, 769 were correctly classified by the Final stage, corresponding to an error-recovery proportion of 0.804. Of the 1567 recordings classified correctly at Base, 91 became incorrect at Final, corresponding to a correct-to-incorrect transition proportion of 0.058. The resulting net difference was 678 additional correct recording-level decisions.

Figure 6 shows the distribution of stable-correct, recovered, degraded, and persistent-error outcomes by activity. Recovery was highest for Walking, with a proportion of 0.935, followed by Mopping/Hoovering at 0.905 and Washing Dishes at 0.841. Brushing Teeth, Washing Face, and Shelving Items showed values of 0.798, 0.755, and 0.754, respectively.

Washing Face therefore showed the largest absolute Base-to-Final accuracy difference but not the highest error-recovery proportion. Its large absolute improvement partly reflected its substantially lower Base-stage accuracy and the greater number of errors available for recovery. Washing Face nevertheless retained the largest proportion of persistent failures and the smallest proportion of stable-correct recordings.

Participant-level Final accuracy is shown in Figure 7. Final performance was generally high for Walking, Mopping/Hoovering, Washing Dishes, Shelving Items, and Brushing Teeth. Greater between-participant variation remained for Washing Face, including two participants whose Washing Face accuracy was below 0.35. The participant-level distribution therefore shows that the high pooled Final accuracy did not correspond to uniform performance across all participant–activity combinations.

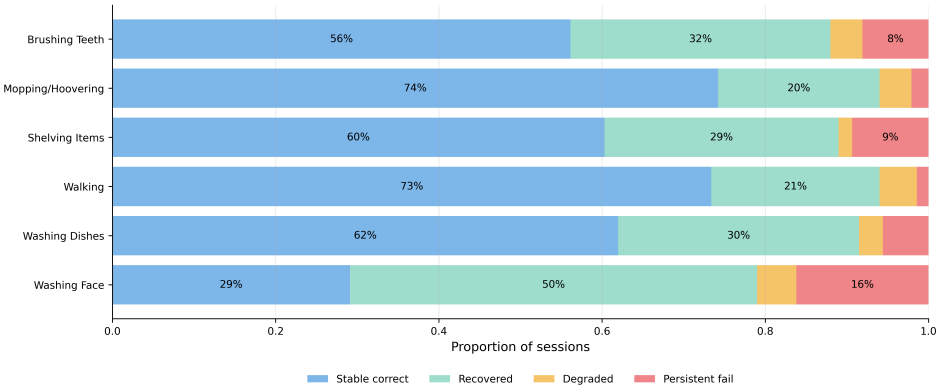


Fig. 6 Per-ADL transition between Base and Final decisions. Recordings are classified as stable correct, recovered (Base incorrect and Final correct), degraded (Base correct and Final incorrect), or persistent error (Base and Final both incorrect).

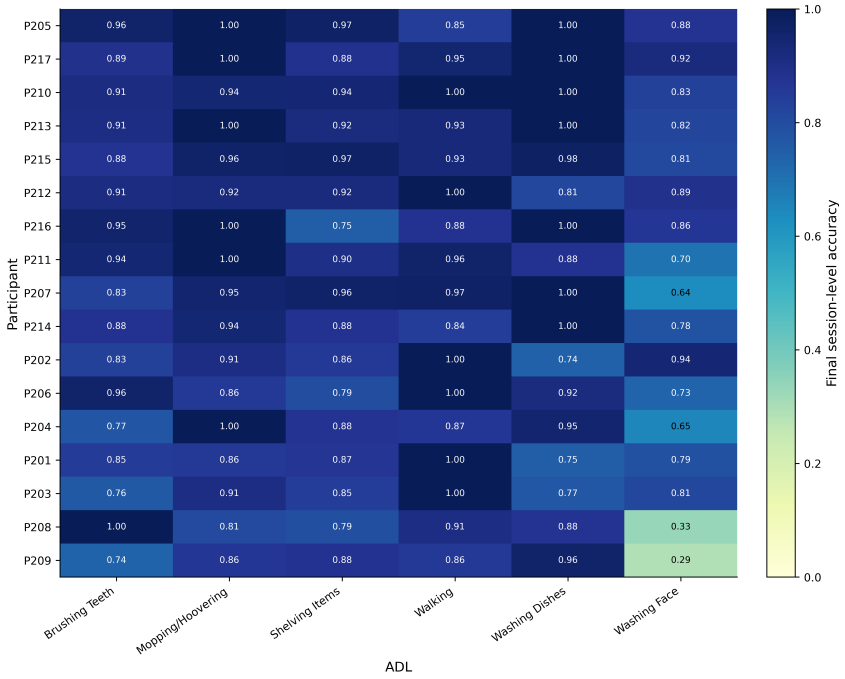


Fig. 7 Participant-wise Final recording-level accuracy for each ADL during Phase 2. Participants are ordered by their mean accuracy across the six activities.

3.6 Activity-specific deployment behaviour and temporal changes

Motion-complexity distributions differed between the simulated-living and free-living recordings in an activity-specific manner (Figure 8). Free-living complexity was higher for Washing Dishes, Shelving Items, and Mopping/Hoovering but lower for Brushing Teeth and Walking. Washing Face showed substantial overlap between the two recording conditions. The transition to free-living use therefore did not produce a uniform increase in calibrated motion complexity across the six ADLs.

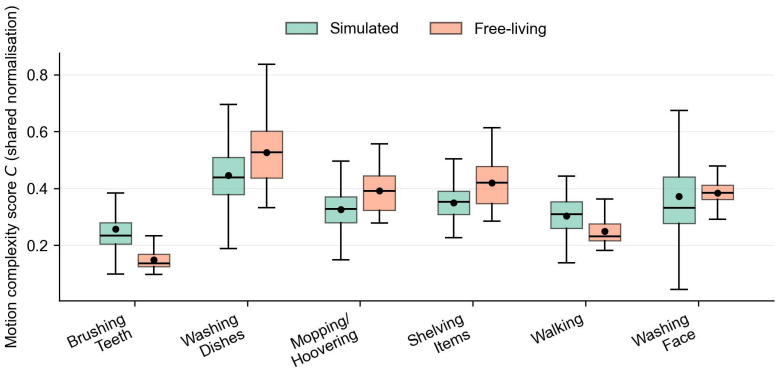


Fig. 8 Calibrated motion-complexity distributions for Phase 1 simulated-living windows and Phase 2 free-living recordings, shown separately for each ADL.

The Final-minus-Base accuracy difference was positive during the first 5-day deployment block for all six ADLs (Figure 9). This early difference is compatible with the immediate availability of the population-level CAM pathway, while participant-specific prototypes could also begin contributing within the same block once their activation threshold had been reached.

Brushing Teeth and Shelving Items showed an increase in the Final-minus-Base difference across later deployment blocks, whereas Walking and Mopping/Hoovering showed smaller and more stable differences. Washing Face showed a larger but more variable difference throughout deployment. Because each 5-day block could contain contributions from both population-level and participant- specific pathways, these temporal patterns were not used to assign causal contributions to individual mechanisms.

The additional Final-stage difference after Global/CAM remained positive across the displayed confirmation-history strata for Brushing Teeth, Shelving Items, Walking, and Washing Dishes (Figure 10). Mopping/Hoovering showed greater variation at lower confirmation counts.

For Washing Face, the Final-minus-Global difference was positive at lower confirmation levels but approached zero in the highest confirmation-history stratum. Increasing confirmation history was therefore not associated with a monotonic increase in additional Final-stage performance for all activities. For Washing Face, this pattern

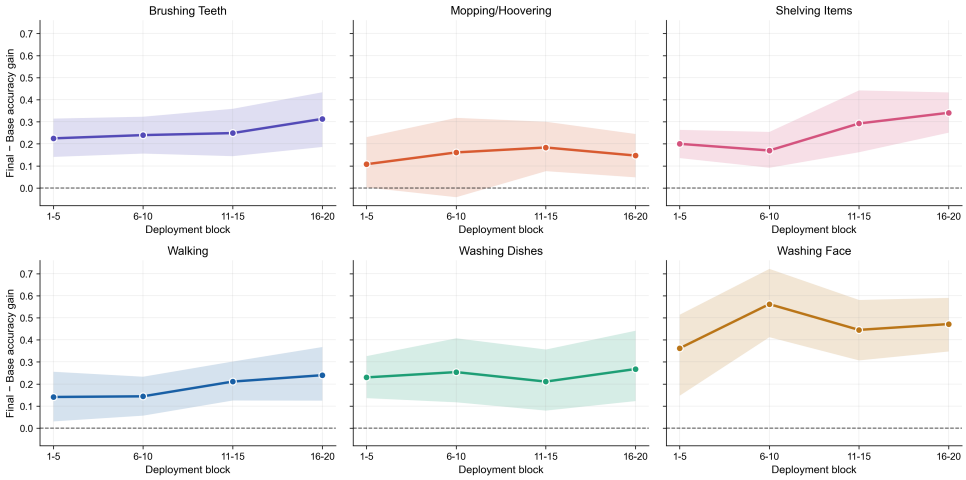


Fig. 9 Final-minus-Base recording-level accuracy difference across four consecutive 5-day Phase 2 deployment blocks. Points show the mean participant-level difference, and shaded regions show participant-clustered 95% confidence intervals.

is consistent with saturation of the available participant-specific benefit or persistent overlap in the learned representation, although the present analysis cannot distinguish these explanations.

Probability-level analysis was consistent with the observed Base-to-Final decision transitions. Recovered recordings showed predominantly positive Final-minus-Base changes in the probability assigned to the participant-confirmed or corrected reference class, whereas degraded recordings were concentrated closer to zero or negative values. Stable-correct recordings showed smaller positive differences, while persistent errors included both weakly positive and negative probability changes (Figure 11).

The distributions were broader for Washing Face and Washing Dishes than for several of the other activities, indicating greater heterogeneity in their stage-wise probability changes.

3.7 Participant-level association with confirmation volume

The association between confirmation volume and participant-level Final-minus-Base accuracy difference was examined separately for each ADL. No statistically significant association was observed.

Brushing Teeth and Shelving Items showed weak negative correlations of $r = -0.21$ and $r = -0.10$, respectively. Mopping/Hoovering, Walking, and Washing Face also showed small correlations, with $|r| \leq 0.15$. Washing Dishes showed the largest positive coefficient, $r = 0.45$, but its confidence interval included zero and the association did not reach conventional statistical significance ($p = 0.072$).

These analyses suggest that confirmation count alone did not explain the magnitude of participant-level adaptive differences. Given the small Phase 2 cohort and the six activity-specific correlations, these findings were treated as exploratory.

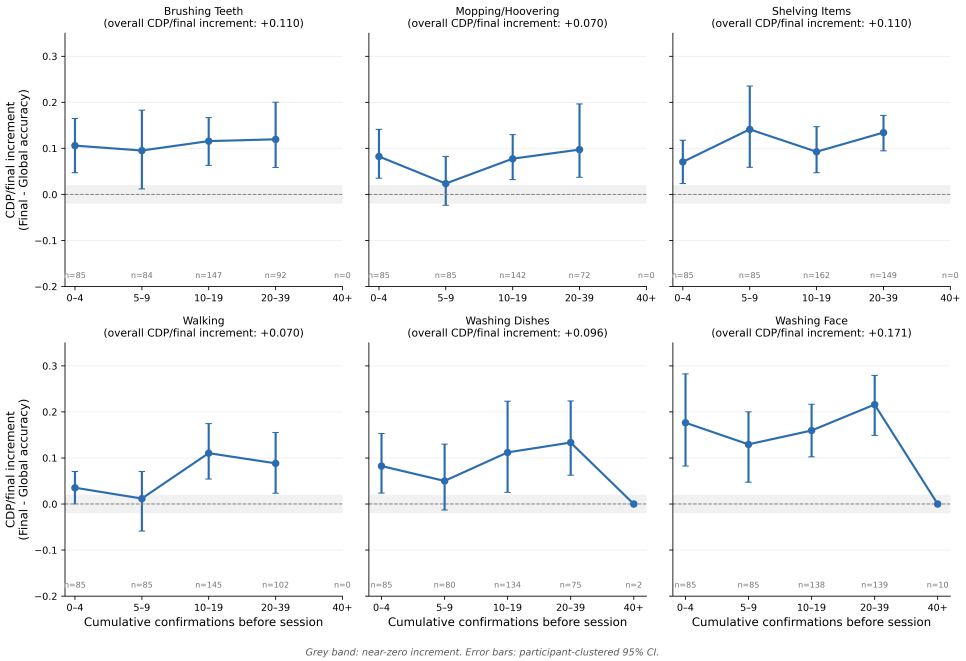


Fig. 10 Final-minus-Global/CAM accuracy difference according to the number of participant confirmations accumulated before each recording. Error bars show participant-clustered 95% confidence intervals, and labels indicate the number of recordings within each confirmation stratum.

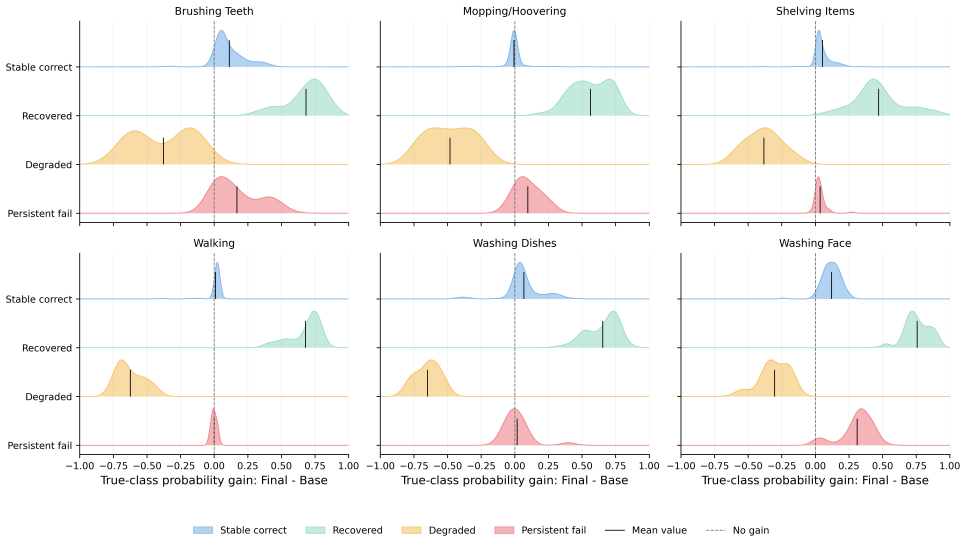


Fig. 11 Distribution of the Final-minus-Base change in reference-class probability across stable-correct, recovered, degraded, and persistent-error recordings. Reference-class probability denotes the probability assigned to the participant-confirmed or corrected deployment reference label.

Table 6 Association between confirmed recordings per ADL and participant-level Final-minus-Base accuracy difference. Confidence intervals were calculated using Fisher's z-transformation.

ADL	r	p	95% CI	Participants with positive difference
Brushing Teeth	−0.21	0.408	[−0.63, 0.30]	100%
Mopping/Hoovering	−0.04	0.891	[−0.51, 0.45]	82%
Shelving Items	−0.10	0.710	[−0.55, 0.40]	100%
Walking	−0.13	0.628	[−0.57, 0.38]	71%
Washing Dishes	0.45	0.072	[−0.04, 0.76]	88%
Washing Face	0.15	0.566	[−0.36, 0.59]	100%

3.8 On-device computational performance

The deployed CoreML model occupied 8.4 MB. Median window-level inference latency was 9.3 ms (IQR, 1.2 ms), and peak process memory was 78 MB. Participant-specific prototype memory required approximately 3.1 kB for six 128-dimensional single-precision prototype vectors.

A benchmarked three-window processing workload required 38 ms for the complete recognition pipeline. This value characterises the specific benchmark workload and should not be interpreted as implying that all Phase 2 recordings contained exactly three analysis windows.

Mean process CPU utilisation during active processing was 22.0%, with a maximum observed utilisation of 68.0% during an individual inference event. The watch remained in the nominal thermal state throughout sustained benchmarking.

Across five controlled one-hour runs, sensing with inference consumed 8.2 ± 0.6 percentage points of battery capacity per hour, compared with 4.7 ± 0.4 percentage points per hour for sensing alone. The incremental difference associated with the recognition workload under this benchmark was therefore 3.5 percentage points per hour.

These measurements characterise computational and power overhead under the controlled benchmark configuration. They should not be interpreted as a direct estimate of unrestricted continuous all-day operation.

4 Discussion

This study examined whether smartwatch-based ADL recognition could respond to the different forms of information that become available during deployment without retraining the learned population representation. Across 2523 participant-initiated free-living recordings from 17 participants, the frozen Base model achieved an accuracy of 0.621, the Global stage after the complete Complexity-Aware Modulation (CAM) pathway achieved 0.784, and the Final stage achieved 0.890. These Phase 2 performance estimates were calculated relative to participant-confirmed or corrected deployment reference labels and were not independently observer-verified. Accordingly, the absolute accuracy values should be interpreted within the human-in-the-loop evaluation procedure used in this study.

Table 7 On-device computational performance of SmartADAPT-Net on an Apple Watch Series 8 running watchOS 10.

Measurement	Assessment procedure	Value
CoreML model size	Model file size on disk	8.4 MB
Window inference latency	500 runs after warm-up; median (IQR)	9.3 ms (1.2 ms)
Benchmarked processing latency	100 repetitions of the three-window processing workload	38 ms
Peak process memory	Xcode Instruments during controlled execution	78 MB
Prototype-memory overhead	Six classes × 128 values × 4 bytes	3.1 kB
Battery consumption: sensing and inference	Five one-hour benchmarks; mean ± SD	8.2±0.6 percentage points/h
Battery consumption: sensing only	Five one-hour benchmarks; mean ± SD	4.7±0.4 percentage points/h
Increment associated with inference	Difference between benchmark means	3.5 percentage points/h
Mean process CPU utilisation	Average across active processing intervals	22.0%
Peak process CPU utilisation	Maximum observed during an inference event	68.0%
Thermal state	Sustained-use monitoring	Nominal

The observed stage-wise pattern was not restricted to pooled accuracy. Macro-F1 increased from 0.635 at Base to 0.793 at Global and 0.892 at Final, while balanced accuracy increased from 0.628 to 0.788 and 0.892, respectively. At the participant level, mean accuracy increased from 0.622 ± 0.071 at Base to 0.887 ± 0.047 at Final, with positive Base-to-Final differences observed for all 17 participants. The smaller between-participant dispersion at Final provides additional evidence that the observed pattern was not confined to a small number of particularly responsive participants.

The principal contribution of these findings is therefore not the Final accuracy alone. Rather, the results illustrate a staged deployment architecture in which an immediately available population-level correction can operate before any participant-specific evidence exists, after which human-confirmed personal information can be incorporated progressively while the neural representation remains fixed.

4.1 From structured model development to free-living deployment

The difference between Phase 1 and Phase 2 Base performance illustrates a familiar challenge in wearable activity recognition. The participant-independent Base model achieved an accuracy of 0.900 under simulated-living evaluation but 0.621 when the frozen population model was applied to the separate Phase 2 cohort during participant-initiated free-living use. Previous work has similarly reported deterioration when

wearable HAR systems developed under structured conditions are applied to unseen users and less controlled environments [9, 18].

The magnitude of the Phase 1 to Phase 2 difference should not, however, be interpreted as a direct estimate of environmental domain shift. The phases used disjoint participant cohorts, differed in their demographic composition, and used different evaluation units: 8-second windows in Phase 1 and recording-level probability aggregation in Phase 2. Free-living recordings additionally allowed greater variation in activity duration, environmental context, object use, and individual execution. The observed reduction in Base performance therefore reflects these factors collectively.

The motion-complexity distributions provide further evidence that the transition was not a uniform shift affecting every activity in the same direction. Some ADLs showed higher median calibrated complexity during Phase 2, while others showed lower values or substantial overlap with the Phase 1 distribution. This supports the broader view that structured-to-deployment mismatch in wearable HAR can be activity-dependent rather than representable by a single global measure of distribution shift.

An important complementary direction is therefore to improve the population representation itself. Recent work on cross-user generalisation, transfer learning, and domain-robust wearable representations seeks to increase performance before deployment adaptation becomes necessary [23, 26, 31]. Representation-level generalisation and the decision-level mechanisms studied here should not be viewed as competing strategies. A stronger population representation could improve cold-start performance, while lightweight deployment-time mechanisms could remain useful for residual user- and context-specific variability that persists after population training.

4.2 Immediate population correction and later personal evidence

The most distinctive pattern in the Phase 2 results was the separation between the information available immediately at deployment and the information accumulated through subsequent user interaction. Overall accuracy increased from 0.621 at Base to 0.784 at Global, an absolute difference of 0.163, before participant-specific prototype evidence was required. Accuracy then increased by a further 0.106 between Global and Final.

Approximately 61% of the observed Base-to-Final accuracy difference was therefore already present at the Global stage. This should be understood as a descriptive property of the staged architecture, rather than as a causal decomposition of the mechanisms. In particular, CAM consists of both the trained residual correction head and the complexity-dependent gate. The present study did not evaluate the trained correction head with the gate removed or fixed to unity, and therefore cannot determine how much of the Base-to-Global difference arose specifically from complexity-dependent gating rather than from the residual correction head more generally.

This distinction does not remove the deployment-level finding. The complete Global pathway required no participant identity, previous confirmations, participant-specific calibration, pseudo-label generation, or deployment-time gradient optimisation. It was available from the beginning of deployment and provided an observation-conditioned population correction before personalised evidence became available. The independent mechanistic contribution of the complexity gate remains an important target for future ablation.

The Phase 1 results also caution against treating population correction as uniformly beneficial. The complexity-aware configuration produced only a modest overall difference under structured evaluation and reduced accuracy for Mopping/Hoovering, despite improving five of the six activity classes. The usefulness of the pathway therefore depended on the activity and evaluation setting rather than reflecting a universal improvement over the Base classifier. This is consistent with the purpose of CAM as a deployment-oriented correction pathway rather than a replacement for a well-performing population classifier.

Confirmation-Driven Personalisation addressed a different stage of deployment. Participant-specific prototypes became available only after sufficient confirmed evidence had accumulated, and their contribution increased gradually through the evidence-adaptive fusion mechanism. The Prototype branch achieved an accuracy of 0.850 on the 2490 prototype-active recordings. Because prototype activation requires prior confirmations, this value characterises a selected post-cold-start subset and should not be interpreted as a directly comparable full-sample estimate against Base or Global performance. The matched Base accuracy on this subset was nevertheless similar to the complete-sample Base value, providing useful descriptive context without eliminating the selection inherent in prototype eligibility.

The confirmation-history analysis further shows that personalisation was not simply a monotonic function of the number of accumulated examples. Final-minus-Global differences remained positive across the displayed strata for several activities, but their magnitude varied and, for Washing Face, approached zero in the highest confirmation-history stratum. Participant-level correlations between confirmation volume and adaptive gain were similarly weak and non-significant for all six ADLs. These results suggest that accumulating more confirmed examples is not sufficient by itself to guarantee progressively larger personalisation benefit. The informativeness and separability of those examples in the frozen embedding space are also likely to matter.

This finding distinguishes the present framework from approaches in which a predefined participant-specific support set is collected before adaptation. Few-shot transfer learning and personalised HAR methods commonly require labelled target-user observations during a calibration or adaptation stage [25, 28]. In SmartADAPT-Net, participant-specific evidence instead accumulates during use through the confirmation interaction. This removes the requirement for a separate participant-specific neural-model training stage, but it introduces a different dependency: personalisation remains contingent on continued user engagement and the reliability of the confirmation labels.

The proposed architecture should consequently be understood as occupying a different deployment configuration rather than as having demonstrated superiority over few-shot, test-time adaptation, or lifelong-learning approaches. For example, lifelong

prototypical methods can update representations or maintain richer support information during streaming operation [33], while optimisation-free test-time adaptation can modify internal normalisation statistics and pseudo-labelled support structures without conventional back-propagation [26]. SmartADAPT-Net instead keeps its neural representation fixed and restricts mutable participant-specific state to compact prototype memory. Direct head-to-head evaluation under equivalent data, label budgets, and hardware constraints would be required to determine comparative performance.

4.3 Combining population and participant-specific evidence

An important feature of the framework is that accumulated participant-specific evidence does not replace the population prediction. The fusion mechanism progressively increases prototype influence while retaining a population contribution throughout deployment. At cold start, the Final and Global distributions are identical because no prototype is active. Once prototypes become available, their maximum contribution is constrained by $\gamma = 0.5$, leaving at least half of the Final weighting with the population pathway.

This conservative design is relevant when interpreting the observed transition behaviour. Of the 956 Base-stage errors, 769 were correct at Final, corresponding to an error-recovery proportion of 0.804. Conversely, 91 of 1567 Base-correct recordings became incorrect at Final, corresponding to a correct-to-incorrect transition proportion of 0.058. The net difference was 678 additional correct recording-level decisions.

The relatively low correct-to-incorrect transition rate is encouraging, but it should not be treated as independent validation of the chosen fusion weight. Prototype influence was deliberately capped at half of the Final decision, which by design limits the degree to which personalised evidence can override the population pathway. The observed 5.8% transition rate therefore reflects the behaviour of this conservative configuration rather than proving that $\gamma = 0.5$ is optimal. Alternative weighting schedules would be required to determine the best trade-off between personalisation strength and preservation of population-level decisions.

More broadly, the results support the conceptual value of retaining both sources of information. A static population model cannot represent every participant-specific execution pattern, whereas an aggressively personalised model may become overly dependent on a small or noisy confirmation history. SmartADAPT-Net treats these sources as complementary: the population branch provides stable support from the outset, and participant-specific information is introduced only as evidence becomes available. The current study demonstrates the feasibility of this staged architecture but does not establish that its particular fusion rule is superior to alternative ways of combining population and personalised predictions.

4.4 Washing Face: correctable systematic error and persistent residual difficulty

Washing Face provides the clearest example of why absolute improvement and residual recognition difficulty need to be considered together. It had the lowest Phase 2 Base

accuracy at 0.339 but showed the largest Base-to-Global difference, reaching 0.619, and the largest absolute Base-to-Final difference, reaching 0.790. At the same time, it remained the lowest-performing activity at Final and had one of the lower error-recovery proportions, at 0.755.

The large absolute improvement therefore does not indicate that Washing Face became the easiest activity to recognise. Its low Base performance created substantially greater headroom and a larger pool of errors that could potentially be corrected. The Base-stage confusion structure also showed that Washing Face errors were concentrated particularly toward Mopping/Hoovering. This suggests that at least part of the Base-stage failure had a systematic decision component that the population correction pathway could partly alter, rather than every error arising from irreducible sensor ambiguity.

The remaining errors tell a different story. Despite the substantial early improvement, Washing Face remained the least accurately recognised activity, showed the highest proportion of persistent failures, and exhibited substantial participant-level variability. Moreover, its additional Final-minus-Global difference approached zero in the highest confirmation-history stratum. Thus, after the more readily correctable component had been reduced, accumulating additional participant-specific evidence did not consistently resolve the remaining errors.

This combination is compatible with two forms of difficulty: systematic population-model error that can be modified at the decision level, and a residual component that is less separable in the available wrist-derived representation. Wrist-worn inertial sensing provides detailed information about arm motion but has more limited access to fine hand actions, object interaction, body context, and environmental information [20, 22]. Activities with partially overlapping wrist trajectories can therefore remain difficult to distinguish even after participant-specific adaptation.

The present study cannot establish that sensor representation alone caused the residual Washing Face errors. Participant variability, unmeasured task context, object interaction, and confirmation error could also contribute. The findings should therefore be described as *consistent with* a limitation of wrist-only inertial representation rather than as proof that Washing Face is fundamentally unrecognisable from the wrist. Multimodal or context-aware sensing is a logical direction for activity pairs that remain poorly separated after both population correction and participant-specific adaptation.

4.5 On-device implementation and deployment feasibility

The implementation results indicate that the adaptive framework can be executed within the computational constraints of the smartwatch used in this study. The CoreML model occupied 8.4 MB, median window-level inference latency was 9.3 ms, complete recording-processing latency was 38 ms under the benchmark procedure, and peak process memory was 78 MB. Participant-specific prototype memory required approximately 3.1 kB. The small mutable state is a direct consequence of retaining the neural model unchanged and updating compact prototype vectors rather than network parameters or a growing replay buffer.

The battery results require more specific interpretation. Under the controlled one-hour benchmark, sensing with recognition consumed 8.2 ± 0.6 percentage points per

hour compared with 4.7 ± 0.4 percentage points per hour for sensing alone, giving an incremental difference of approximately 3.5 percentage points per hour under that benchmark workload.

These measurements support technical feasibility for the participant-initiated operating regime evaluated in Phase 2, in which participants deliberately initiated a limited number of ADL recordings rather than running the recognition pipeline continuously throughout the day. They should not be extrapolated linearly to an 18- or 24-hour operating period, because the benchmark workload does not represent continuous everyday duty cycle. Conversely, the present study does not establish the feasibility of continuous background recognition. Such a system would require separate evaluation and likely additional optimisation of sensing duty cycle, background execution, communication, and power consumption.

4.6 Limitations

The Phase 2 reference labels were obtained after participants viewed the model prediction and should therefore be regarded as participant-confirmed deployment labels rather than independently verified ground truth. This procedure may introduce confirmation or automation bias, and the same feedback was subsequently used to update participant-specific prototype memory. In addition, the Phase 1 and Phase 2 cohorts were disjoint but not demographically matched, and Phase 2 included only 17 participants. The difference between structured and free-living Base performance therefore cannot be attributed exclusively to deployment conditions, and the sample was insufficient for reliable demographic subgroup analyses.

Phase 2 evaluated participant-initiated, closed-set classification of six predefined ADLs rather than continuous open-set recognition, so the findings do not address autonomous segmentation, background activities, transitions, or broader ADL taxonomies. The evaluation also did not isolate the independent contribution of the complexity gate from the residual correction head, and Prototype-stage performance was available only after participant-specific prototypes became active. SmartADAPT-Net was not compared directly with alternative adaptation methods under equivalent conditions, and longer-term studies are needed to assess confirmation burden, prototype stability, and sustained personalisation during extended deployment.

4.7 Implications and future work

The findings suggest several priorities for subsequent evaluation. The most immediate is independent verification of free-living reference labels. Incorporating observer annotation, video-assisted verification where ethically and practically appropriate, or a randomly audited subset of deployment recordings would allow the magnitude of confirmation bias to be estimated rather than merely acknowledged.

A second priority is mechanistic ablation. The residual correction head and complexity gate should be evaluated separately to determine whether the observed Base-to-Global difference depends materially on complexity-conditioned modulation. Alternative complexity descriptors, calibration ranges, and class-conditioned gates could then be assessed only if the gate itself is shown to contribute meaningfully.

Future work should also compare decision-level personalisation with few-shot and continual-learning approaches under equivalent participant-specific evidence budgets. Such comparisons would help determine when a compact frozen-representation strategy is sufficient and when representation-level adaptation provides additional value.

Longer deployments are required to determine whether prototype memory remains stable as participant behaviour changes over time and whether confirmation frequency can be reduced once reliable participant-specific evidence has accumulated. Selective prompting, rather than requesting confirmation after every recording, may reduce interaction burden while preserving the ability to correct drift.

Finally, extension to continuous and open-set recognition will be necessary for less interactive monitoring applications. This will require activity-boundary detection, background-state modelling, out-of-distribution rejection, and evaluation of power consumption under realistic continuous sensing workloads. Multimodal sensing may also be valuable for activities whose residual ambiguity cannot be resolved reliably from wrist inertial data alone.

5 Conclusion

SmartADAPT-Net presents a staged approach to smartwatch-based ADL recognition in which deployment adaptation is separated according to the evidence available at different points in use. A frozen CNN–BiLSTM population representation was combined with an immediately available population-level CAM pathway and a confirmation-driven prototype mechanism that incorporated participant-specific evidence progressively without neural-network retraining.

In a 20-day participant-initiated free-living deployment involving 17 participants and 2523 recordings, accuracy measured against participant-confirmed or corrected deployment reference labels was 0.621 at the Base stage, 0.784 after the complete Global/CAM pathway, and 0.890 at the Final fused decision. Approximately 61% of the observed Base-to-Final accuracy difference was already present at the Global stage, while the subsequent contribution associated with participant-specific evidence varied across activities and confirmation histories. These stage-wise differences describe the behaviour of the complete framework and should not be interpreted as isolated causal effects of the complexity gate or prototype mechanism.

The results also identify important boundaries. Washing Face showed substantial correctable Base-stage error but remained the lowest-performing Final-stage activity, with additional personalisation benefit diminishing at high confirmation history. Phase 2 reference labels were not independently observer-verified, the evaluation was participant-initiated and closed-set rather than continuous, and the independent contribution of the CAM complexity gate was not isolated.

Within these constraints, the study demonstrates the feasibility of combining a stable population representation with lightweight observation-conditioned population correction and progressively accumulated human-confirmed personal evidence on a commercial smartwatch. Future work should establish the contribution of the individual adaptive mechanisms, independently verify deployment labels, compare alternative

adaptation strategies under equivalent conditions, and extend the framework towards longer-term, continuous, and open-set wearable recognition.

Acknowledgements

The authors thank Apple Inc. for providing Apple Watch devices in kind to support the research.

References

[1] Kim, J.-H., Lee, S.B.: Evaluation of activities of daily living: current insights and future horizons. Annals of geriatric medicine and research **29**(2), 143 (2025)

[2] Lavezza, A., Hoyer, E., Friedman, L.A., Daley, K., Steele, A., Rosen, S., Young, D.: Activities of daily living assessment early in hospitalization is associated with key outcomes. American Journal of Occupational Therapy **77**(5), 7705205100 (2023) https://doi.org/10.5014/ajot.2023.050167

[3] Yamada, K., Iwata, K., Yoshimura, Y., Ota, H., Oki, Y., Mitani, Y., Oki, Y., Yamada, Y., Yamamoto, A., Ono, K., *et al.*: Activities of daily living limitation and functional decline during hospitalization predict 180-day readmission and mortality in older patients with pneumonia: A single-center, retrospective cohort study. Respiratory Medicine **234**, 107830 (2024) https://doi.org/10.1016/j.rmed.2024.107830

[4] Khalili, G., Zargoush, M., Huang, K., Ghazalbash, S.: Exploring trajectories of functional decline and recovery among older adults: a data-driven approach. Scientific Reports **14**, 6340 (2024) https://doi.org/10.1038/s41598-024-56606-0

[5] Kwakkel, G., Veerbeek, J.M., Wel, B.C., Wegen, E.E.H., Kollen, B.J.: Diagnostic accuracy of the barthel index for measuring activities of daily living outcome after ischemic hemispheric stroke: Does early poststroke timing of assessment matter? Stroke **42**(2), 342–346 (2011) https://doi.org/10.1161/STROKEAHA.110.599035

[6] Becker, S., Solbrig, S., Michaelis, K., Faust, B., Brockmann, K., Liepelt-Scarfone, I.: Divergence between informant and self-ratings of activities of daily living impairments in parkinson's disease. Frontiers in aging neuroscience **14**, 838674 (2022)

[7] Stull, D.E., Leidy, N.K., Parasuraman, B., Chassany, O.: Optimal recall periods for patient-reported outcomes: challenges and potential solutions. Current Medical Research and Opinion **25**(4), 929–942 (2009) https://doi.org/10.1185/03007990902774765

[8] Ray, S.J., Cherian, J., Liberty, A.M., Hammond, T.A., Shireman, P.K.: Recognition of basic activities of daily living using wearable devices for older adults:

Scoping review. Journal of Medical Internet Research **27**, 67373 (2025)

[9] Ahmed, M.E., Yu, H., Vassallo, M., Koufaki, P.: Advancing real-world applications: A scoping review on emerging wearable technologies for recognizing activities of daily living. Smart Health **36**, 100555 (2025) https://doi.org/10.1016/j.smhl.2025.100555

[10] Cavalcante, A.F., Kunst, V.H.d.L., Chaves, T.d.M., Souza, J.D., Ribeiro, I.M., Quintino, J.P., Silva, F.Q., Santos, A.L., Teichrieb, V., Gama, A.E.F.: Deep learning in the recognition of activities of daily living using smartwatch data. Sensors **23**(17), 7493 (2023)

[11] Ahmed, M.E., Yu, H., Vassallo, M., Koufaki, P.: Enhanced lda with svm and ensemble learning for advanced adl classification using smartwatch data. In: 2024 29th International Conference on Automation and Computing (ICAC), pp. 1–6. IEEE, Piscataway, NJ, USA (2024). https://doi.org/10.1109/ICAC61394.2024.10718845

[12] Zhan, X., Rashid, N., Nemati, E., Ahmed, M.Y., Chandrashekhara, S., Kuang, J.: Mad-fusion: Modality-aware dynamic fusion in identification of activities of daily living. IEEE Internet of Things Journal (2026)

[13] Chan, S., Hang, Y., Tong, C., Acquah, A., Schonfeldt, A., Gershuny, J., Doherty, A.: Capture-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. Scientific Data **11**, 1135 (2024) https://doi.org/10.1038/s41597-024-03960-3

[14] Potter, V., Tran, H., Mobley, D., Bertisch, S.M., John, D., Intille, S.: The physical activity assessment using wearable sensors (paaws) dataset: Labeled laboratory and free-living accelerometer data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **9**(4), 1–32 (2025)

[15] Silva, T.B.C., Cunha Santiago, H., Fernandes, B.J.T.: From controlled benchmarks to unconstrained environments: a pipeline-oriented survey of models, domains, and emerging trends on human activity recognition. Artificial Intelligence Review (2026)

[16] Qureshi, T.S., Shahid, M.H., Farhan, A.A., Alamri, S.: A systematic literature review on human activity recognition using smart devices: advances, challenges, and future directions. Artificial Intelligence Review **58**(9), 276 (2025)

[17] Calatrava-Nicolás, F.M., Miyauchi, S., Rey, V.F., Lukowicz, P., Stoyanov, T., Mozos, O.M.: Embedded inter-subject variability in adversarial learning for inertial sensor-based human activity recognition. In: 2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE, Piscataway, NJ, USA (2025). https://doi.org/10.1109/MLSP62443.2025.11204225

[18] Stojchevska, M., De Brouwer, M., Courteaux, M., Ongenae, F., Van Hoecke, S.: From lab to real world: Assessing the effectiveness of human activity recognition and optimization through personalization. Sensors **23**(10), 4606 (2023) https://doi.org/10.3390/s23104606

[19] Burchard, R., Kirsten, K., Miché, M., Scholl, P., Arnrich, B., Van Laerhoven, K., Lieb, R., Wahl, K.: A real-world dataset for detecting handwashing in daily life using wrist motion data from wearables. Scientific Data (2026)

[20] Proietti, T., Bandini, A.: Wearable technologies for monitoring upper extremity functions during daily life in neurologically impaired individuals. IEEE transactions on neural systems and rehabilitation engineering **32**, 2737–2748 (2024)

[21] Vets, N., Verbeelen, K., Emmerzaal, J., Devoogdt, N., Smeets, A., Van Assche, D., De Baets, L., De Groef, A.: Assessing upper limb functional use in daily life using accelerometry: A systematic review. Gait & Posture **115**, 69–81 (2025) https://doi.org/10.1016/j.gaitpost.2024.11.003

[22] Burchard, R., Brückner, P.-A., Bock, M., Gall, J., Van Laerhoven, K.: Harmes: A multi-modal dataset for wearable human activity recognition with motion, environmental sensing and sound. arXiv preprint arXiv:2605.02596 (2026)

[23] Dhekane, S.G., Ploetz, T.: Transfer learning in sensor-based human activity recognition: A survey. ACM Computing Surveys **57**(8), 1–39 (2025)

[24] Hu, R., Chen, L., Miao, S., Tang, X.: Swl-adapt: An unsupervised domain adaptation model with sample weight learning for cross-user wearable human activity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 6012–6020 (2023). https://doi.org/10.1609/aaai.v37i5.25743

[25] Thukral, M., Haresamudram, H., Ploetz, T.: Cross-domain har: Few-shot transfer learning for human activity recognition. ACM Transactions on Intelligent Systems and Technology **16**(1), 1–35 (2025)

[26] Wang, S., Wang, J., Xi, H., Zhang, B., Zhang, L., Wei, H.: Optimization-free test-time adaptation for cross-person activity recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **7**(4), 1–27 (2024)

[27] Vettoruzzo, A., Bouguelia, M.-R., Rögnvaldsson, T.: Efficient few-shot human activity recognition via meta-learning and data augmentation. Neural Computing and Applications **37**, 19461–19477 (2025) https://doi.org/10.1007/s00521-025-11408-x

[28] Ferrari, A., Micucci, D., Mobilio, M., Napoletano, P.: Deep learning and model personalization in sensor-based human activity recognition. Journal of Reliable Intelligent Environments **9**(1), 27–39 (2023)

[29] Gholamiangonabadi, D., Grolinger, K.: Personalized models for human activity recognition with wearable sensors: deep neural networks and signal processing. Applied Intelligence **53**(5), 6041–6061 (2023)

[30] Celik, S., Bayazit, U.: Gesture detection from semg signals based on similarity learning. Neural Computing and Applications **37**(22), 18663–18677 (2025) https://doi.org/10.1007/s00521-025-11393-1

[31] Kang, H., Hu, Q., Zhang, Q.: Sf-adapter: Computational-efficient source-free domain adaptation for human activity recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **7**(4), 1–23 (2024)

[32] Rey, V.F., Rezende, P.M.B., Zhou, B., Suh, S., Lukowicz, P.: Coa-har: Exploring contrastive online test-time adaptation for wearable sensor-based human activity recognition using sensor data augmentation. Expert Systems with Applications **297**, 129288 (2026) https://doi.org/10.1016/j.eswa.2025.129288

[33] Adaimi, R., Thomaz, E.: Lifelong adaptive machine learning for sensor-based human activity recognition using prototypical networks. Sensors **22**(18), 6881 (2022)