A Personalized Transformer Neural Network for Accurate Recognition of Health-Indicative Complex Activities from Smartphone Sensors

***Abstract*—Continuous monitoring of Activities of Daily Living (ADLs) and Instrumental ADLs (IADLs) of the elderly is vital for their safety, independent living, quality of life and overall health. Declines in the ability to perform these tasks, which require the interplay of musculoskeletal, neurological, and cognitive systems, often indicate underlying health problems. Early detection of such declines can prompt timely interventions. Traditional ADL/IADL assessment was manually done infrequently by a skilled expert. Passive monitoring using data from the built-in sensors of ubiquitous devices such as smartphones offers continuous, objective monitoring of an individual's ADLs in their natural environment. However, the recognition of ADLs from sensor data faces challenges including high intra-class variability due to individualized styles of performing ADLs. One-size-fits-all machine learning models often misinterpret individual nuances in ADL performance, resulting in inaccurate health assessments. Personalization of ADL models can make them robust to inter-subject variability. This paper proposes the Personalized Health Activity Recognition Transformer (P-HART), a personalized transformer-based model that captures temporal relationships in sensor data for robust Complex Activity Recognition (CAR) from smartphone sensor data. In rigorous evaluation on a comprehensive complex ADL dataset, P-HART achieved an F1-score of 92.6% with personalization and 88% without personalization, outperforming baseline models and demonstrating the substantial benefits of ADL model personalization. P-HART facilitates remote health-indicative activity recognition and monitoring.**

***Index Terms*—Activities of Daily Living, Personalized Complex Human Activity Recognition, Smartphone sensors, Transformers**

I. Introduction

Motivation: The aging global population has created unprecedented challenges in healthcare, particularly in continuous monitoring for chronic care management [1]. As of 2022, there were 57.8 million US adults aged 65 and over [2] with 1.3 million of them living in nursing homes and 818,800 in assisted living communities [3]. As an overwhelming majority of older adults prefer to "age in place" rather than relocate to institutional care facilities [4], there is a growing need for effective home-based monitoring solutions that support safe, independent living. ADLs are critical indicators of an individual's functional status and potential to live independently [5]. Since ADLs require coordination between multiple body systems including musculoskeletal, neurological, and cognitive systems, declining performance is often an early warning sign warranting medical intervention and care adjustments. ADLs fall into two categories: basic and instrumental. Basic ADLs such as ambulating, eating, and using the toilet, are required to meet a person's basic needs. Instrumental ADLs (IADLs) include more complex activities necessary for independent living, including meal preparation, managing finances, using transportation, and taking medication. An individual's ability to perform ADLs and IADLs directly correlates with their potential to live safely and independently. ADL and IADL monitoring integratively evaluates physical function, cognitive ability, and psychosocial factors such as motivation to engage in certain activities such as cooking.

Problem: Traditionally, ADL assessments have relied on self- or observer reports [6], [7], which are susceptible to recall bias and inter-rater fluctuations [8], [9]. Additionally, functional decline is often gradual. Initial signs are subtle enough to be overlooked, resulting in missed opportunities for early intervention. This highlights the need for automated monitoring systems. Sensor-based ADL monitoring has emerged but has largely relied on ambient sensors or camera-based systems [10], [11], which require modifications to living environments [12]. Given that over 91% of US adults own smartphones [13], leveraging their built-in sensors for ADL monitoring is a more scalable approach.

Challenges: The recognition of ADLs and IADLs faces technical challenges. Unlike simple activities such as walking that existing Human Activity Recognition (HAR) models can detect, ADLs can be complex, consisting of multiple simple activities that are performed concurrently, or interleaved. *Ranasinghe et al.* [14] defines three types of Complex Activities(CAs): sequential, interleaved and concurrent. Sequential activities consist of multiple simple activities that are performed in a sequence. For example, cooking involves a series of simple activities such as opening the fridge, gathering ingredients and picking up utensils. Concurrent complex activities consist of multiple simple activities performed simultaneously. For example, concurrently cooking while watching the television. Interleaved complex activities consist of a complex activity that is interrupted by a simple activity. For example, an individual may stop cooking to take a phone call before resuming cooking. This complexity limits the efficacy of algorithms that try to match patterns in the training and test sets exactly, transforming Complex Activity Recognition (CAR) into a multi-class, multi-label classification problem requiring sophisticated modeling [15]. Additionally, complex activities exhibit high intra-class and inter-subject variability [16], where the same activity (e.g., "making a sandwich") can be performed in different ways by individuals based on

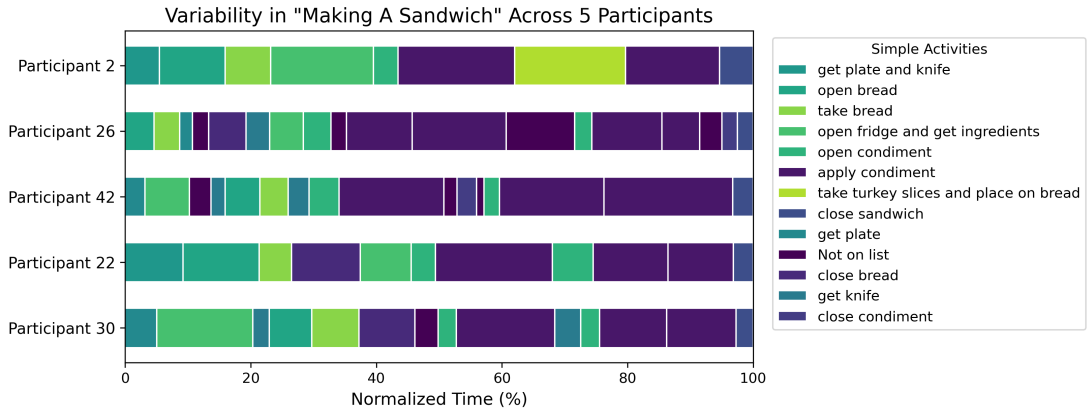

Fig. 1. Illustration of intra-class and inter-subject Variability in an example "Making a Sandwich" ADL

personal style and contextual factors. For instance, Figure 1 illustrates the variability in activity sequences among five randomly selected participants from our dataset, showing that no two individuals performed the exact same sequence of simple activities while making a sandwich. This variability has led to a well-documented challenge of inter-subject generalization in HAR systems [17], where models trained on data from certain individuals often perform poorly on new users due to differences in age, gender, and styles. Personalization is critical for accurate, reliable ADL assessment, particularly in high-stakes domains such as healthcare.

Our approach: We propose Personalized Health Activity Recognition Transformer (P-HART), a custom transformer based architecture with user personalization, for complex HAR using sensor data from a smartphone and simulated smartwatch. P-HART leverages the temporal modeling strengths of transformers while incorporating personalization mechanisms to learn individual performance styles. In rigorous evaluation on our CAR dataset, P-HART achieved a 92.6% F1-score with user personalization vs. 88% without personalization.

Prior work: *Chandrasekaran et al.* [18] introduced CARTMAN, which used topic models to generate features that were classified by a neural network, outperforming AROMA, the previous State-Of-The-Art (SOTA) [19] model for recognizing sequential complex activities. However, CARTMAN did not consider interleaved or concurrent complex activities. More recently, *Ek et al.* [20] proposed HART, a Transformer-based model that achieved SOTA performance in recognizing simple activities but did not consider complex activities.

Contributions:

1) We propose P-HART, a personalized transformer-based architecture to recognize complex ADLs. P-HART learns from fine-grained simple activity labels to understand the composition of complex ADLs from smartphone sensor data.
2) P-HART incorporates a novel multi-task learning strategy that combines a dual-stream cross-modal attention mechanism with a specialized contrastive loss framework (NT-Xent and Temporal Contrastive Loss) to generate robust, context-aware representations of user activity.
3) Rigorous evaluation demonstrating that P-HART significantly outperforms baseline and state-of-the-art models, and quantifies performances gains attributable to personalization. We show that a few-shot adaptation (6-shot) provides the optimal balance for maximizing recognition performance.

II. Related Work

Fine-Grained ADL Recognition: While there have been numerous studies on ADL recognition, not all types of ADLs have been investigated equally. [21] identifies gaps in the literature. For instance, the recognition of ADLs such as bathing and dressing have been under-researched. [22] describes the importance of fine-grained ADL recognition and the use of sensors to capture action-level details that are crucial for assessing subtle changes in an individual's ability to perform ADLs. [23] proposes a method to detect complex ADLs, which first detects atomic activities from wearable sensor data and then uses rank pooling to encode temporal transitions between atomic activities. While their approach improved the accuracy of recognizing sequential complex ADLs, they did not consider concurrent and interleaved ADLs, a limitation that our work addresses.

Smartphone and Smartwatch-Based Recognition: Activities can be recognized from various sensor modalities, including wearable sensors [24], smart home environment [25], Wi-fi [26], audio sensors [27], video [28], or a combination of these modalities. Since we aim to utilize devices already owned by users without modifying the living environment, we focus on smartphones and smartwatches. *Kwapisz et al.* [29] comprehensively demonstrates the viability of using sensors already embedded in smartphones, specifically accelerometers, for recognizing a range of physical activities. *Roy et al.* [30] presents a hybrid approach that augments smartphone-based activity sensing with data from ambient sensors to effectively recognize multiple inhabitants in a home. Their approach discerned individual context by combining person-specific mobile data with person-independent ambient context, significantly improving complex ADL recognition accuracy. *Laput et al.* [31] showcases the ability of commodity smart-

watches to recognize fine-grained hand activities performed in many ADLs (e.g., eating, personal hygiene) by leveraging smartwatch sensors. *Fernández et al.* [32] demonstrates the excellent performance of deep learning architectures in recognizing ADLs from wearable sensor data captured from smartwatches and wearable devices.

III. Dataset

We utilize a previously collected novel CAR dataset consisting of sensor data from 47 participants performing sequential, concurrent and interleaved complex ADLs and IADLs. A paper describing the dataset is in submission. Upon acceptance, the dataset will be publicly available at figshare.com/s/939ec0ab75630c3d1ace. Data were collected from two identical smartphones, mounted on the participant's wrist (simulating a smartwatch) and in their pants pocket. Each device recorded data from five sensors (accelerometer, gyroscope, rotation vector, magnetometer, and gravity) at various sampling frequencies. This yielded 36 raw data columns (18 per device). To ensure consistency, all the sensor data were resampled to 50 Hz and then featurized using sliding windows of varying sizes (1-60 seconds) with different overlap percentages. This yielded a 332-dimensional feature vector. Table I shows the equations for the features.

TABLE I
Formulas used for featurization

Feature	Formulation
Mean	$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i$
Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2}$
Median Absolute Deviation	$\mathrm{median}_i\left(\lvert s_i - \mathrm{median}_j\left(s_j\right)\rvert\right)$
Largest values in array	$\max_i\left(s_i\right)$
Smallest value in array	$\min_i\left(s_i\right)$
Signal magnitude area	$\frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{N}\lvert s_{i,j}\rvert$
Signal Entropy	$\sum_{i=1}^{N}\left(c_i \log\left(c_i\right)\right), c_i = s_i / \sum_{i=1}^{N} s_j$
Interquartile range	$Q3(s) - Q1(s)$
4th order Burg Autoregression	$a = \mathrm{arburg}(s, 4), a \in \mathbb{R}^4$
Pearson Correlation	$C_{1,2}/\sqrt{C_{1,1}C_{2,2}}, C = \mathrm{cov}\left(s_1, s_2\right)$
Freq. signal weighted average	$\sum_{i=1}^{N}\left(is_i\right) / \sum_{j=1}^{N} s_j$
Spectral energy	$\frac{1}{a-b+1}\sum_{i=a}^{b} s_i^2$
Freq. signal Skewness	$\mathrm{E}\left[\frac{(s-\bar{s})^3}{\sigma}\right]$
Frequency signal Kurtosis	$\mathbb{E}\left[(s-\bar{s})^4\right] / \mathrm{E}\left[(s-\bar{s})^2\right]^2$
Largest frequency component	$\arg\max_i\left(s_i\right)$

s: signal vector, N: signal vector length Q: quartile

The study to collect CAR data was conducted in a lab and received Institutional Review Board approval. This dataset contains ≈54 minutes of labeled data per participant, for 29 simple activity and 9 complex activity labels. The participants performed representative ADLs (bathroom activities) and IADLs (cooking activities). The dataset was annotated at two granularities: "sandwich CAs" had fine-grained labels of the constituent simple activities. For privacy reasons, "bathroom CAs" contain only complex-level labels.

IV. Approach

Figure 2 shows the smartphone placement on the user, the list of activities performed and an overview of the classification process. Simple and complex activity labels were one-hot encoded to enable multi-label classification of concurrent simple activities (e.g., talking on the phone while making a sandwich) and interleaved complex activities where multiple tasks can be performed simultaneously.

P-HART (Fig. 3) is a multi-device, transformer-based architecture that uses contrastive loss and user-personalization for enhanced complex ADL recognition. P-HART analyzes time-series sensor data streams from mobile devices in the user's pant pocket and on their wrist. P-HART leverages deep learning to learn robust, personalized representations of complex activities. P-HART uses a series of specialized layers to process two input data streams. Initially, Pocket and Wrist Feature Extractors, each containing a linear layer followed by a ReLU activation layer, extract salient features from raw sensor signals. These features are then passed to a Positional Encoder, which injects data temporal order information, crucial for sequence processing using transformers. The core of each stream is a Transformer Encoder layer that applies multi-headed attention to weight the importance of various elements in the input sequence to capture long-range dependencies in the time-series. To tailor P-HART to individual users, a User Adapter module in each stream fine-tunes and personalizes the learned representations to the user's movement patterns.

A key innovation of the P-HART architecture is the integration of information across the two input streams via cross-modal attention. enabling the learning of complementary information from both sensor streams to enhance complex ADL recognition. The outputs of the cross-modal attention layers are concatenated and processed as follows. First, the output from the cross-modal attention layers is fed into a Recurrent Neural Network (RNN) Encoder suitable for capturing temporal dynamics. The encoder is made up of a bi-directional GRU layer. Second, a Simple Activity (SA) Classification Head analyzes the fused representations to predict the simple activities that make up the complex activity. Since there can be more than one simple activity performed concurrently, this head is trained using Binary Cross Entropy (BCE) loss. The output predictions of the SA classification head is fed into an SA context extractor. The context extractor uses NT-Xent contrastive loss to ensure that the contextual representation of SAs belonging to the same complex activity are closer together while SAs belonging to different complex activities are farther apart. It also uses a temporal contrastive loss to ensure that if two contexts lead to the same activity, their representations are closer while contexts leading to different activities have representations that are far apart.

NT-Xent Loss (Normalized Temperature-Scaled Cross-Entropy Loss): [33] is a type of contrastive loss that maximizes

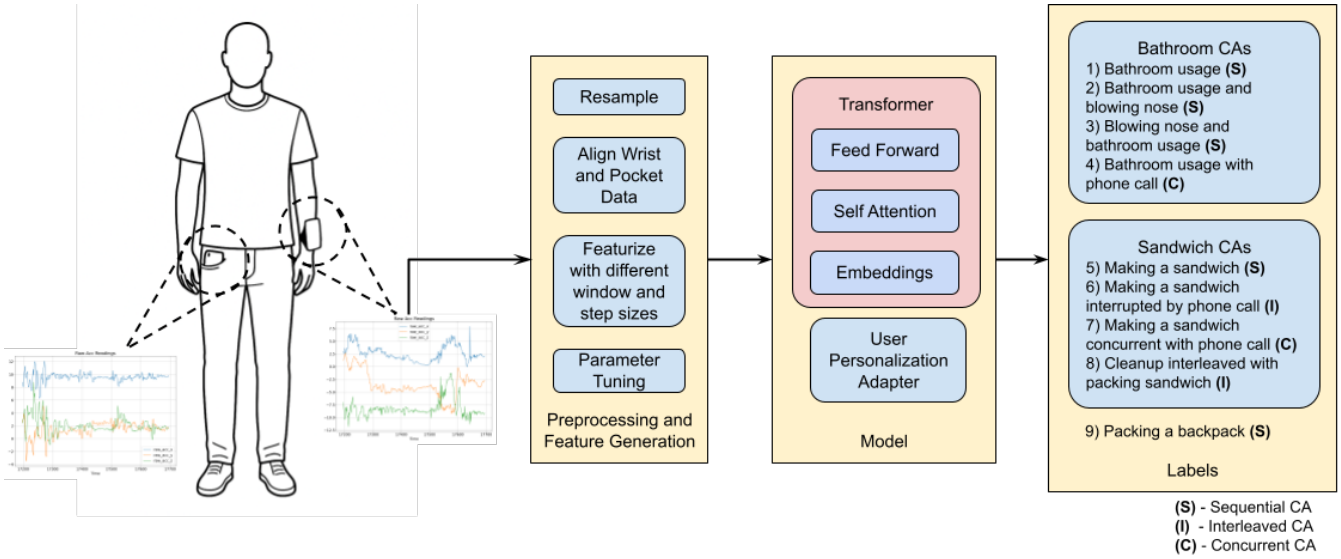

Fig. 2. Overview of our main steps for complex activity classification

the similarity between multiple samples with the same label (positive pairs) while minimizing the similarity of samples with different labels (negative pairs) in a batch. To calculate the NT-Xent loss, all feature vectors are first normalized to unit length. This simplifies the similarity calculation to a matrix multiplication, as the dot product of two unit vectors is their cosine similarity, $\hat{z}_i = \frac{z_i}{\|z_i\|_2}$. Then the similarity matrix is computed. It contains cosine similarities between every pair of feature vectors in the batch and scales them by a temperature parameter (τ). A lower temperature makes the distribution of similarities sharper, increasing the penalty for dissimilar items being close, expressed as $\text{sim}(z_i, z_j) = \frac{\hat{z}_i \cdot \hat{z}_j^T}{\tau}$. A key part of this implementation is how it handles one-hot encoded labels, especially for interleaved activities where a sample can have multiple 1s. By performing a matrix multiplication of the label with its transpose, a label similarity matrix is created. If two samples share any common activity, their dot product will be at least 1. It is made to be a binary mask, where 1 if they are a positive pair and 0 otherwise. For a single positive pair of samples (i,j), the loss is given by 1

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j))}{\sum_{k=1, k\neq i}^{N} \exp(\text{sim}(z_i, z_k))} \tag{1}$$

Temporal Contrastive Loss: function applies contrastive learning principles to the temporal dimension of sequential data. The goal is to teach the model to represent and understand activity transitions such that if two different temporal contexts (pairs of consecutive activity embeddings) lead to the same next activity, their representations are similar. Conversely, if they lead to different next activities, their representations are dissimilar. Two linear layers are used to project the temporal context. Let z_{t-1} and z_t be the embeddings at time $t-1$ and t respectively. Their concatenation is $c_t = [z_{t-1}; z_t]$. $h_1 = \text{ReLU}(c_t W_1^T + b_1)$ and $p_t = h_1 W_2^T + b_2$. The embedding is L2 normalized and the embedding $z_{b,t}$ for batch b at time step t. $z_{b,t,d}$ is the d-th component of this embedding. The normalized embedding is denoted as $\hat{z}_{b,t}$.

$$\hat{z}_{b,t,d} = \frac{z_{b,t,d}}{\sqrt{\sum_{d'=1}^{\text{embedding\_dim}} (z_{b,t,d'})^2}} \tag{2}$$

The normalized embedding of the previous step ($\hat{z}_{t-1}$) and the current step ($\hat{z}_t$) are concatenated to form the input to the context projection. $c_t = [\hat{z}_{t-1}; \hat{z}_t]$ Then the temperature-scaled cosine similarity between all pairs of flattened projected contexts, is computed. Let p'_k and p'_l be the k-th and l-th vectors in the projected contexts p_t after normalization and flattening. $\text{S}_{k,l} = \frac{p'_k \cdot (p'_l)^T}{\tau}$ where τ is the temperature parameter. The label similarity matrix is computed in which two contexts p'_k and p'_l are "positive" pairs based on their target next activity labels. Let L'_k and L'_l be the one-hot encoded target label vectors for contexts p'_k and p'_l respectively.

$$\text{LabelSim}_{k,l} = \min(\max((L'_k \times (L'_l)^T), 0), 1) \tag{3}$$

$$\mathcal{L}_k = -\log \left(\frac{\sum_{l=1, l\neq k, \text{LabelSim}_{k,l}=1}^{N'} \exp(\text{S}_{k,l})}{\sum_{l=1, l\neq k}^{N'} \exp(\text{S}_{k,l})} + \epsilon \right) \tag{4}$$

where, ϵ is a small constant for numerical stability. The temporal contrastive loss is the mean of $\mathcal{L}_k$ over all contexts k that have at least one positive pair. Finally, the output of the cross-modal attentions layers, RNN Encoder, and SA context extractor are concatenated and used in the CA head, to classify complex activities. The training of this last head is supervised by a composite loss function that includes the SA (BCE), CA (BCE), NT-Xent, and Temporal Contrastive Losses.

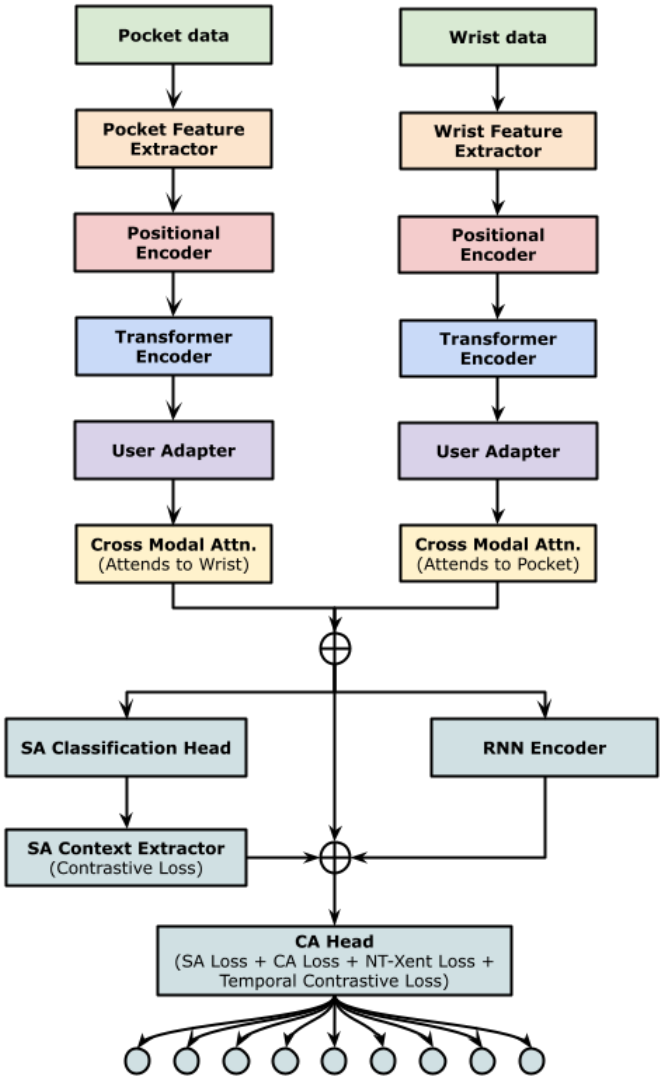

Fig. 3. Personalized Transformer-based CAR Model Architecture

V. Evaluation

P-HART is compared to baseline activity recognition models using 5-fold cross-validation with subject-wise splitting. A 70%:15%:15% training:validation:test split ratio was utilized. Averages of metrics Accuracy, F1-Score, Precision, and Recall are reported.

We compared against the following traditional machine learning baseline models 1) *Support Vector Machine (SVM)* [34] with a Radial Basis Function (RBF) kernel, 2) *Random Forest (RF)* [35], and 3) *XGBoost* [36]. We also compared against the following deep learning models.

Multilayer Perceptron (MLP) with 4 hidden layers each containing 256, 128, 64 and 32 units respectively. The final layer is a sigmoid layer with 9 outputs. A dropout layer was used for regularization. The Adam optimizer with weight decay was used. The optimizer used a BCE loss function with custom weights for each class. The class weights were calculated using normalized inverse class frequency $f_i = \frac{n_i}{N}$, where f_i is the frequency of class i, n_i is the number of samples in class i, and N is the total number of samples. $w_i^{\text{norm}} = \frac{w_i}{\sum_{j=1}^{C} w_j} \times C$ where w_i^{norm} is the normalized weight for class i, and C is the total number of classes.

DeepConvLSTM [37] has performed well in prior simple HAR work. We used a DeepConvLSTM model with 1 CNN layer followed by 1 LSTM layer and a fully connected sigmoid layer. Adam was used as the optimizer with weight decay. A BCE loss function was used with custom per-class weights calculated using normalized inverse class frequency.

CARTMAN [18] first uses a Latent Dirichlet Allocation (LDA) topic model to generate sensor features, in order to discover the simple activities ("topics") that constitute the larger complex activity. These topic-based features are classified using DeepConvLSTM with self-attention.

HART [20] leverages the Transformer's attention mechanism to weight the importance of various time-points in time series sensor readings. This captures long-range dependencies and complex temporal patterns more effectively than RNNs. Minimal changes were made to the HART model to accommodate the multi-label classification.

P-HART implementation Weights and Biases' sweep functionality [38] was used to determine optimal values of hyperparameters including the number of layers, the dimensions of layers and the personalization adapter, number of samples and number of adaptation steps used for personalization.

We conducted comprehensive model evaluation including comparison to baselines and an ablation study to quantify the contribution of individual model components. To assess the contributions of user personalization, P-HART model performance with and without personalization were compared. For the personalized P-HART model, two key parameters were investigated: (1) minimum number of test instances per user required for effective adaptation, and (2) optimal number of adaptation steps. Various few-shot learning scenarios were evaluated, where zero-shot indicates that no user-specific test data was used for adaptation, one-shot employs a single test instance, and so forth. After systematic experimentation, 6-shot adaptation was determined to achieve the optimal balance between performance improvement and minimal user-specific data for personalization.

To evaluate the impact of personalization on individual users, we conducted a Leave-One-Participant-Out (LOPO) cross-validation experiment. As shown in Figure 4, user personalization consistently improved performance for 41 out of 47 participants. The six participants who did not benefit from personalization already had high performance, suggesting limited room for improvement from personalization. Notably, participant 5 exhibited near-zero performance without personalization, indicating potential data quality issues. Hence, participant 5 was excluded from all subsequent experiments, resulting in a final dataset of 46 participants. During evaluation, various regularization techniques were employed including data augmentation, dropout, and label smoothing. We randomly selected 1 of 5 data augmentation functions: jitter,

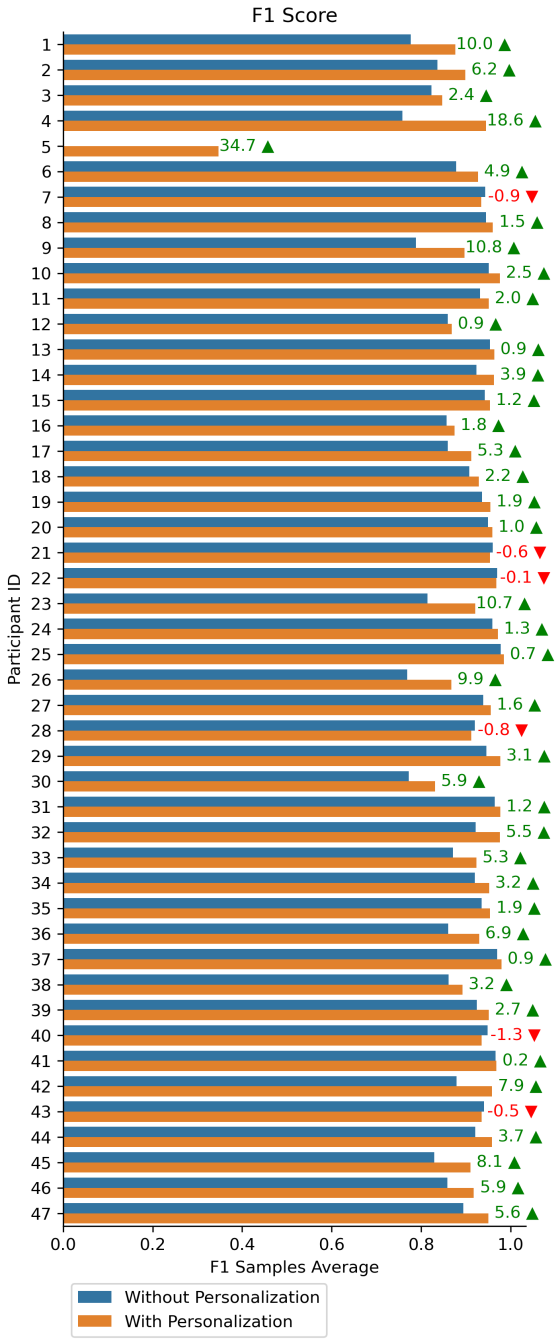

Fig. 4. Per-Participant F1-Score With and Without Personalization

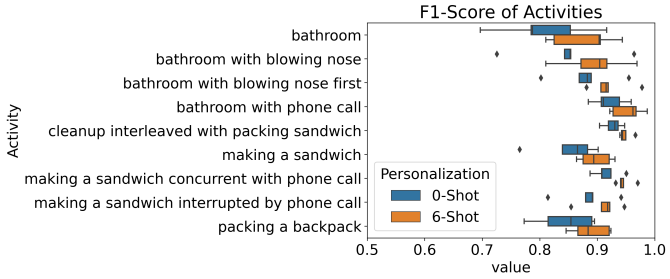

Fig. 5. Impact of Personalization of F1-Score of Activities

TABLE II
COMPARISON WITH BASELINES

Model	Exact Accuracy	F1-Score	Precision	Recall
Machine Learning Models				
SVM	0.000±0.00	0.235±0.03	0.147±0.02	0.600±0.07
RF	0.391±0.05	0.392±0.05	0.393±0.05	0.392±0.05
XGBoost	0.569±0.04	0.600±0.04	0.601±0.04	0.606±0.04
Deep Learning Models				
MLP	0.468±0.07	0.514±0.06	0.512±0.06	0.528±0.06
DeepConvLSTM	0.385±0.06	0.404±0.06	0.404±0.06	0.409±0.05
CARTMAN	0.262±0.09	0.283±0.09	0.279±0.09	0.291±0.09
HART	0.673±0.06	0.693±0.07	0.712±0.06	0.692±0.07
P-HART (Ours)	**0.912±0.02**	**0.926±0.02**	**0.927±0.02**	**0.927±0.02**

VI. DISCUSSION

Our findings show the significant contributions of personalization for complex activity recognition, with P-HART achieving a 92.6% F1-score, surpassing the 88% F1-score of its non-personalized version and baseline models. Personalization effectively addressed inter-subject variability in ADL/IADL performance caused by individual performance styles that generalized models may misinterpret. Few-shot personalization improved performance for most users and demonstrated that our Transformer backbone with a contrastive learning strategy is highly effective for this task. While this is a promising result, personalization offered minimal gains for users who already have high performance with the generalized model and slightly reduced the performance for 6 participants. These 6 participants had high baseline scores without personalization, indicating their ADL performance style was captured well using the generalized model. We speculate that personalization introduced minor inaccuracies by overfitting to a non-representative adaptation set. Furthermore, the P-HART's sensitivity to data quality was highlighted by the need to exclude one participant with anomalous sensor readings. This suggests that a deployed system must handle such real-world issues. The evaluation also revealed a dependency on a small, labeled adaptation set, as zero-shot personalization was insufficient to achieve the peak performance possible with the few-shot model.

The limitations of this work mainly stem from the absence of fine-grained labels for the "bathroom CAs". In future work these missing labels can be handled using transfer learning. Other future directions include exploring unsupervised personalization, extending the P-HART to more activities and cohorts, and conducting in-the-wild studies.

scaling, rotation, time warping and random dropout. Table II shows the overall performance of our personalized P-HART model compared to baselines. To assess the contribution of each sensor modality, Table III details the per-activity F1-scores for models trained using only pocket data, only wrist data, and both data streams combined. Figure 5, demonstrates the effectiveness of personalization across different activities using F1-score. Table IV shows the results from an ablation study, revealing that each of P-HART's components contribute non-trivially to its performance.

TABLE III
F1-SCORES PER ACTIVITY - PER PHONE

Activity	Pocket Phone	Wrist Phone	Both Phones
Sequential CA			
bathroom	0.877±0.06	0.845±0.08	0.878±0.06
bathroom with blowing nose	0.862±0.07	0.829±0.08	0.895±0.06
bathroom with blowing nose first	0.889±0.06	0.840±0.07	0.920±0.04
making a sandwich	0.876±0.02	0.837±0.04	0.897±0.03
packing a backpack	0.872±0.02	0.854±0.07	0.888±0.03
Interleaved CA			
cleanup interleaved with packing sandwich	0.934±0.01	0.942±0.01	0.948±0.01
making a sandwich interrupted by phone call	0.883±0.04	0.739±0.10	0.910±0.03
Concurrent CA			
bathroom with phone call	0.948±0.04	0.884±0.06	0.953±0.03
making a sandwich concurrent with phone call	0.926±0.03	0.704±0.09	0.946±0.01

TABLE IV
ABLATION STUDY

Model	Exact Accuracy	F1-Score	Precision	Recall
Our Model	0.912±0.02	0.926±0.02	0.927±0.02	0.927±0.02
Without User Adapters	0.858±0.03	0.878±0.03	0.879±0.03	0.882±0.03
Without Cross Modal Attention	0.906±0.02	0.921±0.02	0.921±0.02	0.925±0.02
Without Simple Activity Branch	0.886±0.02	0.901±0.02	0.903±0.02	0.903±0.02
Without RNN Encoder	0.882±0.02	0.898±0.02	0.899±0.02	0.902±0.02

VII. CONCLUSION

To address the limitations of generalized machine learning models in recognizing real-world complex activities, this paper introduced P-HART, a personalized multi-device Transformer that learns individual ADL and IADL performance styles from smartphone sensors to mitigate inter-subject and intra-class variability in activity data. The personalized P-HART model achieved a 92.6% F1-score, significantly outperforming the non-personalized version (88% F1-score) and baseline models, demonstrating the contributions of personalization. This research advances robust health-indicative complex activity recognition, paving the way for accurate remote health monitoring to support independent living and timely interventions.

REFERENCES

[1] A. Bookman and D. Kimbrel, "Families and elder care in the twenty-first century," *The Future of Children*, pp. 117–140, 2011.

[2] U. C. Bureau, "Nat. population projections tables: Main series," 2023.

[3] C. N. C. for Health Statistics, "National post-acute and long-term care study," 2022.

[4] M. Toth *et al.*, "Trends in the use of residential settings among older adults," *J. Gerontology: Series B*, vol. 77, no. 2, pp. 424–428, 2022.

[5] P. F. Edemekong, D. Bomgaars, S. Sukumaran, and S. B. Levy, "Activities of daily living," 2019.

[6] S. Cloutier *et al.*, "Trajectories of decline on instrumental activities of daily living prior to dementia in persons with mild cognitive impairment," *Int'l J. Geriatric Psychiatry*, vol. 36, no. 2, pp. 314–323, 2021.

[7] M. A. Dubbelman *et al.*, "Decline in cognitively complex everyday activities accelerates along the alzheimer's disease continuum," *Alzheimer's Research & Therapy*, vol. 12, pp. 1–11, 2020.

[8] R. E. Ready, B. R. Ott, and J. Grace, "Validity of informant reports about ad and mci patients' memory," *Alzheimer Disease & Associated Disorders*, vol. 18, no. 1, pp. 11–16, 2004.

[9] K. Persson, A. Brækhus, G. Selbæk, Ø. Kirkevold, and K. Engedal, "Burden of care and patient's neuropsychiatric symptoms influence carer's evaluation of cognitive impairment," *Dementia and geriatric cognitive disorders*, vol. 40, no. 5-6, pp. 256–267, 2015.

[10] A. Chan *et al.*, "Evidence and user considerations of home health monitoring for older adults: scoping review," *JMIR aging*, vol. 5, no. 4, p. e40079, 2022.

[11] R. Ullah *et al.*, "Vision-based activity recognition for unobtrusive monitoring of the elderly in care settings," *Technologies*, vol. 13, no. 5, p. 184, 2025.

[12] R. Fritz, D. Cook, *et al.*, "Detecting older adults' behavior changes during adverse external events using ambient sensing: Longitudinal observational study," *JMIR nursing*, vol. 8, no. 1, p. e69052, 2025.

[13] Pew Research Center, "Mobile fact sheet," tech. rep., Pew Research Center, Washington, D.C., Nov. 2024.

[14] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity rec. systems with regard to performance and evaluation," *Int'l J. Distrib. Sensor Nets*, vol. 12, no. 8, p. 1550147716665520, 2016.

[15] S. Dernbach *et al.*, "Simple and complex activity recognition through smart phones," in *Int'l Conf. Intell. Env.*, pp. 214–221, IEEE, 2012.

[16] Y. Xu *et al.*, "Learning multi-level features for sensor-based human action recognition," *Perv. & Mobile Comp.*, vol. 40, pp. 324–338, 2017.

[17] D. Xiong, S. Wang, L. Zhang, *et al.*, "Generalizable sensor-based activity rec. via categorical concept invariant learning," 2024.

[18] K. Chandrasekaran *et al.*, "Cartman: Complex activity recognition using topic models for feature generation from wearable sensor data," in *Int'l Conf. Smart Computing (SMARTCOMP)*, pp. 39–46, IEEE, 2021.

[19] L. Peng, L. Chen, Z. Ye, and Y. Zhang, "Aroma: A deep multi-task learning based simple and complex human activity rec. method using wearable sensors," *Proc. ACM IMWUT*, vol. 2, no. 2, pp. 1–16, 2018.

[20] S. Ek, F. Portet, and P. Lalanda, "Transformer-based models to deal with heterogeneous environments in human activity recognition," *Personal and Ubiquitous Computing*, vol. 27, no. 6, pp. 2267–2280, 2023.

[21] S. J. Ray, J. Cherian, A. M. Liberty, T. A. Hammond, and P. K. Shireman, "Recognition of basic activities of daily living using wearable devices for older adults: Scoping review," *JMIR*, vol. 27, p. e67373, 2025.

[22] S. Pan, M. Berges, J. Rodakowski, P. Zhang, and H. Y. Noh, "Fine-grained activity of daily living (adl) recognition through heterogeneous sensing systems with complementary spatiotemporal characteristics," *Frontiers in Built Environment*, vol. 6, p. 560497, 2020.

[23] M. A. Nisar, K. Shirahama, F. Li, X. Huang, and M. Grzegorzek, "Rank pooling approach for wearable sensor-based adls recognition," *Sensors*, vol. 20, no. 12, p. 3463, 2020.

[24] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[25] M. Stikic, T. Huynh, K. Van Laerhoven, and B. Schiele, "Adl recognition based on the combination of rfid and accelerometer sensing," in *Proc. Pervasive Health*, pp. 258–263, IEEE, 2008.

[26] S. Tan, L. Zhang, Z. Wang, and J. Yang, "Multitrack: Multi-user tracking and activity rec. using commodity wifi," in *Proc. CHI*, pp. 1–12, 2019.

[27] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *Proc. ACM IMWUT*, vol. 3, no. 1, pp. 1–18, 2019.

[28] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, "Recognition of activities of daily living with egocentric vision: A review," *Sensors*, vol. 16, no. 1, p. 72, 2016.

[29] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[30] N. Roy, A. Misra, and D. Cook, "Ambient and smartphone sensor assisted adl recognition in multi-inhabitant smart environments," *J. ambient intelligence and humanized computing*, vol. 7, pp. 1–19, 2016.

[31] G. Laput and C. Harrison, "Sensing fine-grained hand activity with smartwatches," in *Proc. CHI*, pp. 1–13, 2019.

[32] A. D. R. Fernández, D. R. Fernández, M. G. Jaén, J. M. Cortell-Tormo, *et al.*, "Recognition of daily activities in adults with wearable inertial sensors: Deep learning methods study," *JMIR Medical Informatics*, vol. 12, no. 1, p. e57097, 2024.

[33] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, pp. 1597–1607, PmLR, 2020.

[34] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[36] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, pp. 785–794, 2016.

[37] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[38] L. Biewald, "Experiment tracking with weights and biases," 2020. Software available from wandb.com.