This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3586074

A Feature-Augmented Transformer Model to Recognize Functional Activities from in-the-wild Smartwatch Data

Bryan Minor, Colin Greeley, Ryan Holder, Brian Thomas, Lawrence B. Holder, and Diane J. Cook, *Fellow, IEEE*

***Abstract*—Human activity recognition (HAR) from wearable sensor data traditionally identifies atomic movements (e.g., sit, stand, walk). However, many medical fields require recognizing functional activities—higher-level, goal-directed behaviors (e.g., errands, socialize, work). Functional activity recognition is critical for cognitive health assessment, rehabilitation, post-surgical recovery, and chronic disease management, yet remains largely unexplored due to its inherent complexity and variability for in-the-wild settings.**

This work addresses these challenges by investigating methods for functional HAR and introducing a novel approach that augments feature representations with feature token-transformer embeddings to improve classification performance. We compare a range of machine learning and deep learning methods, analyzing their ability to generalize across a diverse population. Additionally, we present ArWISE, a large-scale functional activity dataset collected longitudinally from n=503 participants, consisting of over 32 million labeled points. Our experiments demonstrate the advantages of incorporating feature embeddings into functional HAR models, particularly in handling real-world variability and data sparsity. By bridging the gap between atomic movement recognition and functional behavior modeling, this work lays the foundation for more advanced, behavior-aware applications in digital health and human-centered AI.

***Index Terms*— activities of daily living, functional activities, human activity recognition, ubiquitous computing, wearable sensor data**

I. Introduction

HUMAN activity recognition (HAR) provides a structured vocabulary for describing sensor-derived behavior patterns. Activity-labeled data enable a host of health benefits for behavior tracking, detecting deviations from typical routines, and developing behavior-aware interventions for health and well-being. By assigning labels to time series data collected from wearable sensors, HAR facilitates the characterization of activities of interest, supports personalized health monitoring, and allows for population-level comparisons in research studies. As wearable technology becomes increasingly integrated into everyday life, accurate HAR systems play a vital role in applications ranging from fitness tracking to clinical decision-making.

Traditional HAR research has primarily focused on recognizing *atomic movements*—short, discrete physical actions such as walking, sitting, standing, or climbing stairs. While valuable, these atomic movements provide only a low-level representation of human behavior. In contrast, *functional activities*—which include work, hobbies, hygiene, socialization, running errands, eating, housework, and sleep—encompass higher-order, goal-oriented behaviors that span extended time periods and often involve multiple physical movements in combination. Recognizing functional activities presents unique challenges. Unlike atomic movements, which are typically constrained to simple biomechanical motions, functional activities are composite in nature, involving a sequence of diverse actions and interactions with the environment. They can also be semantically ambiguous, because different individuals may perform the same functional activity in highly variable ways. For instance, "working" may involve sitting at a desk and typing for one person, while for another, it may involve standing and engaging in manual labor. Additionally, functional activities exhibit substantial human variability, influenced by individual habits, contextual factors, and personal preferences. These challenges make it difficult to develop robust models for functional HAR using conventional motion-sensor-based recognition approaches. Furthermore, for solutions to be usable outside of controlled settings, they must handle the inconsistences of data collected in the wild, such as activity interruptions, collection inconsistencies, and missing data.

In this work, we investigate methods for functional HAR and propose a hybrid approach that combines a traditional classifier with feature token-transformer embeddings. Our approach leverages the power of transformer-based representations to capture complex temporal dependencies

*This work was supported in part by the National Institutes of Health under grants R35AG071451 and R01AG065218 and by the National Science Foundation under grants 1954372 and 2240615.

Bryan Minor is with Washington State University, Pullman, WA 99164 USA (e-mail: bminor@wsu.edu).

Colin Greeley is with Washington State University, Pullman, WA 99164 USA (e-mail: colin.greeley@wsu.edu).

Ryan Holder is with the University of California at Irvine, Irvine, CA 92697 USA (e-mail: rholder@uci.edu).

Brian Thomas is with Washington State University, Pullman, WA 99164 USA (e-mail: bthomas1@wsu.edu).

Lawrence B. Holder is with Washington State University, Pullman, WA 99164 USA (e-mail: holder@wsu.edu).

Diane J. Cook is with Washington State University, Pullman, WA 99164 USA (e-mail: djcook@wsu.edu).

within functional activities, improving classification performance. Additionally, we introduce a large-scale functional HAR dataset, called ArWISE, consisting of over 37 million labeled points collected longitudinally from 503 participants in real-world settings. The methods and dataset described in this paper are publicly available and provide an unprecedented opportunity for advancing functional HAR research in diverse, in-the-wild scenarios.

II. Related Work

Human activity recognition (HAR) from wearable sensor data is now a standard focus for biomedical informatics, machine learning, and time series research. HAR methods are evolving because the task faces challenges when analyzing data from noisy sensors and diverse individuals in real-world settings. Many existing efforts are designed to process data collected in controlled settings [1], yet recent advances apply to more complex settings as well. Many recent approaches rely on deep networks to automate representation learning [2], though some researchers observe that traditional random forests are sometimes more effective [3]. Recent work also successfully combines deep learning with classical methods to model related problems in biomedicine using a hybrid approach [4], [5].

When working with raw time series data, attention mechanisms improve performance by focusing on parts of the input such as selected features, time points, and channels [6], [7], [8], [9]. Researchers further enhance performance by pretraining the network with available information such as the participant, device types, and sensor positions [10], [11]. If data are available from multiple sensor modalities, these can be fused to improve model robustness [12], [13], [14], [15]. Researchers are finding that large language models can play a beneficial role in fusing heterogeneous data sources [16], [17], [18], [19].

Recognizing activities from continuous data collected in the wild is a daunting task. Some of the challenges include activity variability as well as distribution shifts between persons and over time [20]. Unlike laboratory datasets, real-world data are not pre-segmented, activities may co-occur, and there is a scarcity of reliable ground-truth labels. The conditions under which data are collected vary so dramatically between studies that proposed evaluation metrics encourage leave-one-subject-out and leave-one-dataset-out cross validation [21], [22]. Dai et al. [10] incorporate domain-invariant contrastive learning to account for some of these variances, and Su et al. [23] include user information as features then employ adversarial disentanglement to remove irrelevant information.

Another approach to tackling real-world variance is to incorporate transfer learning, or domain adaptation, into the pipeline. Thukral et al. [24] utilize a teacher-student self-trainer to align multiple datasets, Mazankiewicz et al. [25] align feature distributions across users, and Wang et al. [11] define normalization steps that are resilient to distribution shifts between users. Wilson et al. [26] learn domain-invariant features to adapt labeled data from multiple users to a new person with no ground-truth labels.

Laboratory data simplify the recognition task because each sequence corresponds to one activity and the location is fixed. In real-world settings, recognition is improved when transitions are detected [27], [28] or learned jointly with the activity models [29], [30], [31]. Alazeb et al. [32] further noted the value of jointly localizing and recognizing activities.

Obtaining ground-truth activity labels typically relies on individuals labeling their own tasks, but this increases user burden and interrupts the monitored activities. As a result, labeled real-world are extremely rare [33], [34], [35]. Time series augmentation [13], [36], [37], [38], [39], [40] and synthetic data creation [15], [41] help when more data, or more balanced classes, are required. Other researchers make strategic use of unlabeled data to pretrain a deep network [42] or self-train pseudo-labeled training data [21], [43], [44].

This work uniquely positions itself by considering the problem of modeling and recognizing high-level functional activities from continuous smartwatch data. Furthermore, we examine the ability of deep learning methods to handle large amounts of diverse time series data collected from hundreds of participants across multiple study cohorts[1].

III. ArWISE Dataset

Wearable data are a primary driver of current research in digital medicine and behavioral health and medicine. However, the maturity of machine learning approaches is severely constrained by the lack of large-scale, diverse labeled data. The dearth of such data is a common lament in the field and impedes progress in developing robust activity models that generalize to real-world scenarios.

We introduce ArWISE (Activity recognition from in-the-Wild SmartwatchEs)[2], a dataset containing labeled and unlabeled data collected by Apple Watches. ArWISE represents readings collected from 20 studies in 2 countries over 8 years.

A. Data Collection

Data collection followed a consistent protocol for each cohort. The process was reviewed and approved by the Institutional Review Board at Washington State University (protocol number 14460). To participate, participants needed to understand English and sign an informed consent form. Participants were given an Apple Watch to wear each day on their non-dominant arm. While they wore the watch, a custom app collected 3D accelerometer and gyroscope readings at 10Hz. Additionally, the app collected the person's location every minute or when the magnitude of the acceleration vector exceeded a threshold.

At random times throughout each day, the smartwatch prompted the participant to select an activity from a scroll-down list that best described their current activity. The distribution of user-provided labels across 12 activity categories is shown in

[1] The dataset, software, and pretrained model are made available to the community at github.com/WSU-CASAS/ArWISE.

[2] The ArWISE dataset is available at datadryad.org/share/N0QT27E71qLeb1nOcqYR5cP-mevflxke7T3us4BKZtM.

Fig. 1. The label was applied to five minutes of sensor readings ending at the time of the participant's response.

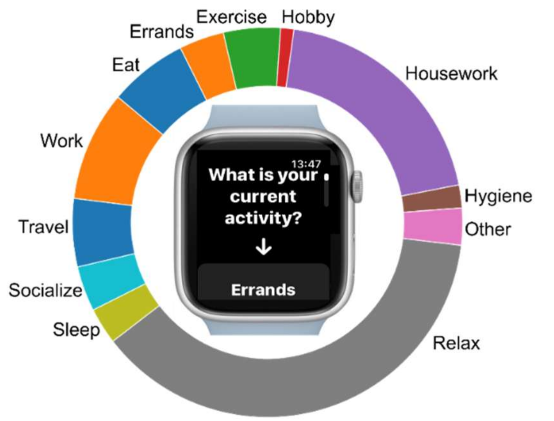

Figure 1. Distribution of data in ArWISE. The chart shows the distribution of labeled points across activity categories. The numbers of samples in each category are: Eat=2,675,397; Errands=1,755,729; Exercise=1,839,335; Hobby=501,637; Housework=8,360,616; Hygiene =752,455; Other=1,504,911; Relax=15,634,352; Sleep=1,337,699; Socialize=1,504,911; Travel=2,173,760; Work=3,762,277.

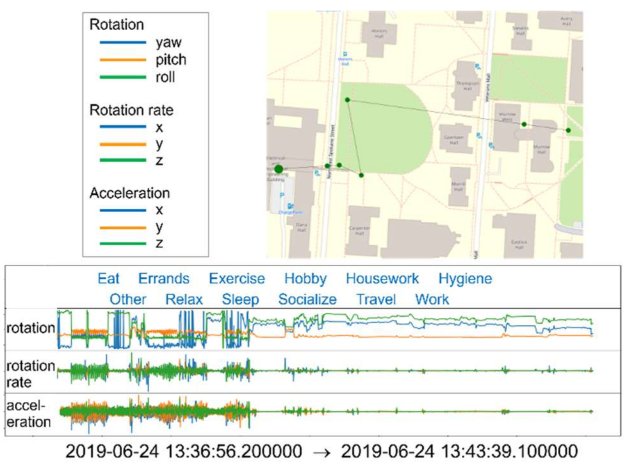

Figure 2. The data annotation tool visualized date, time, movement, and location. The user could move forward, backward, and zoom in and out of timeframes, labeling selected timeframes with one of the specified activity categories.

Additionally, an external annotator provided labels for a much greater density of data collected for selected participants in cohorts 7, 17, and 18. This person used a tool, shown in Fig. 2, that visualized 3D movement data, a map of visited locations, and time stamps, at arbitrary time frames.

While the data collection mechanism was the same for all study cohorts, other parameters varied. These include the number of participants, participant demographics, length of data collection, and other clinical variables that were collected. A summary of study cohort parameters is given in Table I.

B. Dataset Characteristics

The ArWISE dataset is unique among the resources that are typically available for human activity recognition. Some of the most-analyzed datasets reflect movement categories based on data that are collected in controlled settings [45], [46]. However, more recent wearable sensor datasets represent activities observed in uncontrolled settings. Although 150 participants are monitored for only 24 hours with movement-only sensors, Capture-24 [47] includes labels for functional activities of household chores, sports, and sleep in real-world settings. ExtraSensory [48] monitors a smaller set of 60 participants with up to 20 seconds of movement and location readings but provides diverse activity and location. The UK Biobank [49] offers 7 days of accelerometry data for 100,000+ participants and Intuition [50] longitudinally observes 23,004 participants, though no ground-truth labels are provided for these data.

The ArWISE dataset contains 41,803,079 labeled points from 503 participants across 15 cohorts and 469,881,358 total points for 854 participants across 20 cohorts. Each point represents one minute of data. ArWISE offers unique benefits for HAR analysis, including a large set of participants, functional activity labels, longitudinal observations, and consistency in the data collection mechanism.

TABLE I
ARWISE COHORTS, LISTED CHRONOLOGICALLY.

Cohort	Sample	Points	Study/participant characteristics
1	4	$4.44{\times}10^{6}$	Younger adults, self-reported activities
2	185	$1.22{\times}10^{8}$	HOA/SCD/MCI[a], English and Spanish self-reported activities
3	56	$7.08{\times}10^{7}$	Younger adults, no activity labels
4	46	$3.03{\times}10^{7}$	HOA/SCD/MCI, self-reported activities
5	10	$1.27{\times}10^{7}$	Older adult pairs, no activity labels
6	35	$1.38{\times}10^{7}$	HOA/SCD/MCI, no activity labels
7	37	$8.48{\times}10^{6}$	HOA/SCD/MCI, self-reported activities and expert-annotated activities
8	9	$1.53{\times}10^{5}$	Younger adults, self-reported activities
9	15	$1.00{\times}10^{7}$	Younger adults, self-reported activities
10	13	$1.41{\times}10^{7}$	Younger adults, self-reported activities
11	3	$1.95{\times}10^{6}$	Younger adults, self-reported activities
12	18	$1.83{\times}10^{7}$	Younger adults, self-reported activities
13	10	$7.87{\times}10^{6}$	Younger adults, self-reported activities
14	22	$1.70{\times}10^{6}$	Younger adults, self-reported activities
15	21	$3.56{\times}10^{5}$	HOA/SCD/MCI, no activity labels
16	6	$3.25{\times}10^{6}$	Younger adults, self-reported activities
17	103	$1.74{\times}10^{8}$	HOA/SCD/MCI, self-reported activities and expert-annotated activities
18	16	$2.28{\times}10^{7}$	HOA/SCD/MCI, self-reported activities and expert-annotated activities
19	16	$2.25{\times}10^{7}$	HOA/SCD/MCI, self-reported activities
20	229	$8.00{\times}10^{8}$	HOA/SCD/MCI, no activity labels

[a]HOA=healthy older adult, SCD=subjective cognitive decline, MCI=mild cognitive impairment

C. Data Preprocessing

Our functional activity recognition models consider both raw time series data and engineered features. Table II summarizes the features that are available for both cases.

TABLE II
ARWISE RAW AND ENGINEERED DATA FEATURES.

	Category	Feature
Raw (10Hz)	time	date and time
	motion	yaw, pitch, roll, rotation rate (x,y,z), acceleration (x,y,z)
	location	latitude, longitude, altitude, course, speed
Eng-ineered (1 min)	time	time of day (radians, sin, cos), day of week
	motion	mean & stdev (each raw movement variable), mean & stdev (rotation vector magnitude, acceleration vector magnitude)
	location	mean & stdev (course, speed) mean & stdev (distance from home, latitude distance from home, longitude distance from home) mode & stdev (bearing from home)
Activity label	eat, errands, exercise, hobby, housework, hygiene, relax, sleep, socialize, travel, work, other	

We imputed missing values (with mode for location and median for other features) and dropped data points where there was not a complete minute of sensor readings leading up to the label. We also applied z-score normalization to each feature.

For the engineered features, we aggregated values over one minute leading up to the user (or expert) label. Time of day was encoded using two sinusoidal features that reflect its periodicity: $\sin(2\pi t/86400)$ and $\cos(2\pi t/86400)$, where t is the number of seconds past midnight. We did not use raw location values to preserve user privacy and because the values do not generalize between individuals. Instead, we defined a person's home as the location visited most often at the beginning of each day. We then extracted the Haversine distance and trigonometric bearing from the person's home location.

D. Functional Activity Challenges

Recognizing functional activities from a large dataset poses several unique challenges. First, data representing hundreds of individuals in their everyday lives are highly variable. Models built for one set of individuals may not generalize to a new person. Second, self-reporting leads to a label sparsity that can impact performance.

Third, functional activities introduce semantic ambiguity. To illustrate the point, consider the UMAPs in Fig. 3. The first plot shows a random sampling of points that belong to the Relax and Work activities. As the plot indicates, most of the Work points form a dense and relatively continuous cluster. In contrast, Relax points are much more dispersed and partially embedded in the Work category, which may create difficulties in predicting these classes. This overlap is intuitive, because tasks that comprise Work for some individuals (e.g., sitting at a computer) represent relaxation for others and vice versa. In contrast, Exercise and Sleep, shown in the second UMAP, are more clearly differentiated. In both cases, the classes have a core region but many additional isolated clusters.

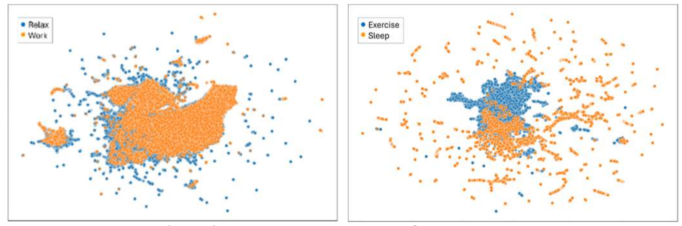

Figure 3. 2D UMAP of data samples from (left) the Relax and Work categories and (right) the Exercise and Sleep categories, described using engineered features from Table II.

IV. METHODS

Our goal is to design a machine learning model that can recognize functional activities from many data points, activity categories, and participants. The models summarized in Section II are not directly applicable because they are designed for different modalities, activity types, and experimental conditions. However, we include models that represent prior methods. We additionally assess the impact of supervised and unsupervised model pretraining, and transformers. All deep networks utilize an Adam optimizer, ReLU activation, sparse categorical cross-entropy loss, an accuracy metric for self-validation, and a softmax classifier output layer with 12 classes. Hyperparameters for baseline methods are consistent with those used by these structures in HAR. For the FT+RF method, we perform grid search over key hyperparameters. Additional fine-tuning had minimal impact on relative performance.

For all classifiers, we include data augmentation to address class imbalance by creating synthetic points via jitter and feature-wise permutation. Approximately 5 synthetic points are created for every real point, though the actual number of synthetic points for each class is inversely proportional to the relative class size. We also added self-supervised training to the methods by predicting activities for the unlabeled data points and refining the model using the original training data combined with points whose label confidence exceeded 0.8. Because these enhancements improved predictive accuracy by $\leq$0.1%, we do not report the results of these steps separately for each model.

A. Time Series Networks

Many human activity recognition approaches extract features from raw time series using deep networks. We consider five such methods, each of which processes ArWISE data windows that contain 100 continuous time steps. The first is a 1D CNN, an approach that is a popular baseline for recognition of atomic movements due to its ability to capture local temporal patterns [38], [51]. The second is a long- and short-term time series network (LSTNet [52]) that combines a CNN component, a recurrent component, a Skip-RNN, an autoregressive component, and a fusion layer. Third, we include a transformer method that is adapted to handle time series data (TST [53]).

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3586074

Fourth, we consider a model that combines Time2Vec time series embeddings with a transformer (T2T). The final model is a transformer-based method (TimesNet [54]) that discovers multiple periodic time series patterns.

B. Tabular Data Networks

Time series models often exert costly computational demands. However, other deep networks have been explored that efficiently process tabular data such as the engineered features shown in Table II. In response to these challenges, we also consider tabular deep networks.

We start by considering a fully-connected deep network (DNN) consisting of two dense layers with 30% dropout. We also include a transformer-based model for tabular data (TabTransformer [55]) that processes categorical features along with the continuous features, by embedding each categorical feature into a dense vector space. We further boosted this baseline model using additional normalization, residual connections, and dropout.

C. DNN with Supervised Contrastive Pretraining

To boost the DNN tabular network, we introduce supervised contrastive pretraining. Here, ground truth labels are accessed by a contrastive loss function to group similar activity points together in the learned representation space. The learned embeddings are fine-tuned by the classification deep network described in the previous section.

This embedding model is a feedforward network containing dense layers of sizes 128, 64, and 64 embedding units. The network also uses 30% dropout, and an L2 normalization layer. The contrastive loss computes a similarity matrix between embedding points using cosine similarity divided by a temperature to control focus on hard negatives (dissimilar examples / different classes) versus easier positives (similar examples / same class). The temperature parameter controls the focus on hard negatives versus easier positives.

D. DNN with Unsupervised Autoencoder Pretraining

While classifiers need labeled data to train the model, some methods utilize sampled unlabeled data to pretrain a model. This pretraining enables the network to create a representation that reflects the entire dataset. We design a masked autoencoder to act as an unsupervised pretrainer.

Using an autoencoder, the pretrainer learns to reconstruct its own input after randomly-selected features have been masked. The result of this process is that the pretrained model learns a representation that captures dependencies between the features. In our case, the pretrainer selects 30% of the input features for masking. The autoencoder is a feedforward network with two dense layers (128 and 64 nodes) interspersed with 30% dropout layers. The network uses a mean squared error loss function to measure reconstruction quality as the squared difference between the original and reconstructed values. As in the previous model, this pretrainer is later fine-tuned by the deep network described in Section IV.B for activity classification.

E. Random Forest

Random forest classifiers are popular for many predictive clinical tasks. Recently, in some cases they have been outshined by deep networks that process large numerical datasets to automatically extract features and learn nonlinear relationships between feature values to predict class values. In our functional HAR task, we consider random forests because of their ability to process large noisy datasets and stability when handling imbalanced data. We start with a random forest that processes the engineered features listed in Table II with 100 trees, a Gini impurity measure, and instance weights that are inversely proportional to relative class frequency.

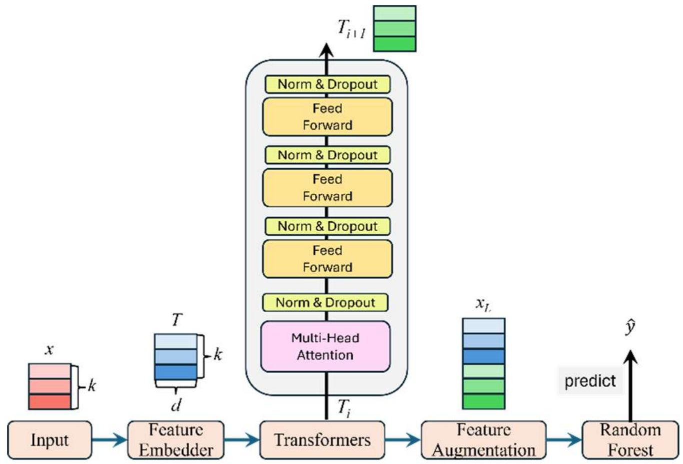

Figure 4. FT-Transformer-Augmented Random Forest architecture.

A notable difference between ArWISE and prior benchmark datasets summarized in Section III is the inclusion of location features. When learning atomic movement types, motion features have been sufficient to model and discriminate between classes. We hypothesize that for functional activities, time and location will also be important. To validate this hypothesis, we report performance of the random forest using only 3D motion features and compare the results with random forest models that use all features listed in Table II, including motion, location, and time.

F. Random Forest + FT-Transformer Embeddings

Transformer models enhance deep networks by employing self-attention to capture influences between features. Typically, these influences are relationships between tokens in a sequence. Harnessing the power of deep neural networks for heterogeneous tabular data is more challenging because relationships must be extracted that do not rely on inherent spatial or sequential neighbor information.

Unlike traditional transformers, recent methods apply self-attention at the feature level. We first capture dependencies between ArWISE engineered features using an FT-Transformer (Feature Tokenizer + Transformer) [56]. This model, shown in Figure 4, converts each data point x described by a feature vector of size k into an embedding matrix T of shape $k \times d$, which it feeds through L transformers. The feature embedder applies element-wise multiplication between a learnable weight vector and the feature and adds a bias term, projecting the feature into a d-dimensional space. Each transformer layer consists of a multi-head self-attention component to learn dependencies between features, feedforward layers, and normalization with

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3586074

dropout to promote model stability and prevent overfit.

A grid search through hyperparameters finalized the FT-transformer architecture. This model contains an embedder with dimension d=64, a Dense layer of 64 nodes, 4 transformer layers with 30% dropout and 64-node feedforward, and a mean squared error loss function to promote self-supervised feature reconstruction. Pretraining is performed using sampled unlabeled data from all ArWISE cohorts.

Next, instead of relying solely on the model embeddings, we merge the engineered features from Table II with the corresponding embeddings to create the FT-augmented feature vector x_L, which is input to the random forest for functional activity classification. We hypothesize that combining a deep transformer embedder with a classical model will boost performance. Embeddings introduce features that are not modeled manually, the random forest is robust to noisy data, and the hybrid method combines different inductive biases to improve generalization over either method alone.

V. Experimental Results

We evaluate the performance of the classification methods on the ArWISE data. We randomly extract training and test points from the global pool dataset with an 80/20 split. A stride of 1 is used within training data, but training and test data points do not occur within the same minute, avoiding the resulting bias due to temporal data leakage. Results are averaged over five training/test splits. We trained the models on a machine with an Intel Core i9-13900K 5.8GHz CPU, a RTX 4090 24GB GPU, and 192GB RAM.

We assess performance using four metrics: Accuracy, F1 score, Matthews correlation coefficient, and Top-3 accuracy. As Fig. 3 demonstrates, the activity categories reflect some overlap. As a result, we report top-3 accuracy as the proportion of points where the correct activity class appears among the model's three highest-probability predictions, useful in settings where multiple plausible labels may exist and should be handled accordingly. Performance for the classifiers using these metrics is summarized in Table III.

A. Time Series Models vs Tabular Data Models

Models that process 3d motion data for recognizing atomic movements frequently process raw sensor readings as a time series. Deep networks such as 1D CNNs, RNNs, or LSTMs are effective at discovering temporal features but may suffer when processing high-dimensional inputs. As Table III shows, the time series models perform consistently well, with MCCs in the 0.37-0.42 range. TimesNet slightly edges out the others for F1 and MCC, though LSTNet demonstrates the best accuracy. Performance is better than random guess (F1≅.08, MCC≅0.00) and the results validate that temporal modeling is important for functional HAR.

The DNN tabular model performs on par or better than most of the time series models, especially in MCC. On the other hand, TabTransformer struggles across all metrics. Among the most effective methods are DNN with pretraining that include supervised contrastive loss and unsupervised autoencoder. The results indicate that simple dense networks can capture salient information from these data. Furthermore, representation learning helps, even for shallow networks with engineered features.

TABLE III

HAR results on ArWISE dataset. Models include time series classifiers (CNN, LSTNet, TST, TimesNet, T2T), tabular deep networks (DNN, Trasnformer), Dnn enhanced with supervised contrastive pretraining (Dnn+Cl) and masked autoencoder pretraining (Dnn+MAP), Random forest with motion sensor features only (RF Motion), all manual features (RF All), Random forest using FT-Transformer embeddings (RF Embed), and Random forest using engineered features combined with embeddings (RF+FT). The top performer in each category of models is hghlighted in italic font, and the Top overall performer is highlighted in bold font.

Model	Accuracy	F1	MCC	Top3
CNN	0.568	0.541	0.369	*0.809*
LSTNet	*0.619*	0.566	0.413	0.808
TST	0.608	0.556	0.408	0.794
T2Vec+Transformer	0.610	0.566	0.421	0.797
TimesNet	0.609	*0.574*	*0.423*	0.780
DNN	0.608	0.549	0.483	*0.829*
TabTransformer	0.376	0.206	0.000	0.664
DNN + CL	*0.616*	0.549	*0.492*	0.826
DNN + MAP	0.610	*0.551*	0.486	0.827
RF Motion	0.612	0.545	0.488	0.814
RF All	0.666	0.649	0.578	0.807
RF Embed	0.675	0.647	0.568	0.863
RF + FT	**0.777**	**0.761**	**0.712**	**0.928**

B. Inclusion of Location and Time Features

Many HAR methods model activities solely based on movement (accelerometer and gyroscope) readings. For functional activities in the wild, we hypothesize that these temporal features need to be supplemented with additional context such as location and time of day. Including these additional features in the model represents a fundamental component of functional HAR in the wild. While atomic movements are reproducible at multiple times of day and in multiple locations, functional activities are often further characterized by their location (e.g., travel occurs outside the home and location may change within a single occurrence) and time of day (e.g., sleep most often occurs at night).

To validate our hypothesis, we assessed the impact of including location and time features for the same subset of 50 participants. We evaluated the performance of the random forest classifier using only motion features in comparison with a random forest that uses all features from Table II. As Table III illustrates, the inclusion of these context features improves performance for all metrics, increasing F1 by 19%, MCC by 18%, and rivaling the performance of DNN+CL with all features.

C. Hybrid Models

Next, we consider the models that incorporate FT-

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3586074

Transformer embeddings. Using a feature vector comprised solely of embeddings improved accuracy and MCC, though F1 dropped slightly. This result indicates that embeddings capture some patterns and relationships not represented by the manually engineered features.

The hybrid model that combines the random forest with the feature transformer embeddings performs best among all considered models by a wide margin, especially in MCC and top-3 accuracy. Based on MCC, TimesNet performs best among time series classifiers (MCC=0.423), DNN with pretraining performs best among the tabular deep learners (MCC=0.492), and random forest performs best overall (MCC=0.578 for engineered features, MCC=0.712 for FT-augmented features). These results offer validation that feature engineering combined with embedding is effective for functional HAR.

D. Leave-One-Out Cross Validation

We evaluate a subset of models (the CNN time series model, DNN tabular model, and RF+FT model) for their ability to generalize predictive performance to new participants. Here, we focus on a subset of participants. Most participants do not have examples of all 12 activity classes, which complicates assessment of leave-one-out cross validation. To address this issue, we selected all participants who had labeled instances for activities categories that included, but were not limited to, the six activity categories found with most participants: Eat, Errands, Exercise, Housework, Hygiene, and Work.

Table V summarizes the results for the 45 participants from 6 cohorts using a CNN model, DNN model, and RF+FT model. We compare results for three evaluation strategies: 1) within-subject training and testing, 2) pooled training and testing across all data points with temporal separation, and 3) leave-one-subject-out cross validation.

TABLE IV

FUNCTIONAL HAR RESULTS FOR N=45 PARTICIPANTS CONTAINING LABELED EXAMPLES FOR 6 ACTIVITY CATEGORIES. THE COMPARED MODELS ARE A DEEP NEURAL NETWORK (DNN) AND A RANDOM FOREST WITH FT-TRANSFORMER AUGMENTED FEATURES (RF+FT). EVALUATION IS CONDUCTED USING DATA FROM THE ENTIRE SET OF PARTICIPANTS (ALL), TRAINING AND TESTING WITHIN EACH PERSON SEPARATELY (IND), AND LEAVE-ONE-OUT CROSS VALIDATION (LOOCV). FOR THE IND AND LOOCV CASES, MEAN RESULTS ARE REPORTED.

Model	Method	Accuracy	F1	MCC	Top-3
CNN	All	0.499	0.447	0.348	0.662
	Ind	0.484	0.465	0.392	0.660
	Loocv	0.407	0.384	0.239	0.744
DNN	All	0.688	0.652	0.533	0.929
	Ind	0.624	0.600	0.429	0.862
	Loocv	0.400	0.382	0.184	0.749
RF + FT	All	0.806	0.791	0.720	0.971
	Ind	0.739	0.726	0.621	0.918
	Loocv	0.523	0.492	0.318	0.833

Unsurprisingly, the results in Table V reveal that the models consistently perform better when processing data from all participants rather than a leave-one-out evaluation. The nature of functional activities varies between individuals, so the availability of even a small amount of training data for each person boosts predictive performance. Because some of the participants did not offer many labeled instances, the results of training and testing separately on each individual are lower than training a model on all participants. In this situation, the RF + FT model consistently outperforms the DNN, for all metrics and all evaluation conditions.

Finally, we built a confusion wheel to visualize the types of errors that are made by the RF+FT model for the functional HAR task. The wheel, shown in Figure 5, shows the distribution of true/false positives and negatives for each activity class. The thickness of the edge between two nodes is proportional to the relative number of times the corresponding activities are confused for each other. We see from the confusion wheel that Relax is a very heterogeneous category. The activity is often confused with Work, Eat, and Housework. Sleep and Relax also have a thick connection. Travel does not share much connection with other categories except Errands, and Exercise is distinct from most categories except Hobby and Relax. These observations show the intuitive overlap between functional categories and highlight the challenges in predicting and tracking functional activities.

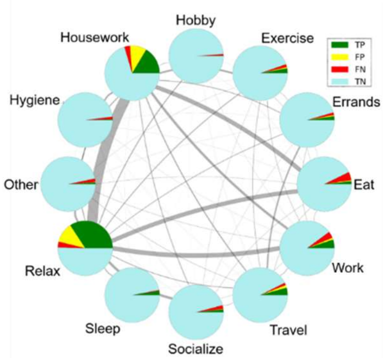

Figure 5. Confusion wheel. Each node shows a distribution of points for true positives / negatives and false positives / negatives. Nodes are connected by an edge when a point belongs to one node (activity) and is predicted as the other. The thickness of the edge indicates the relative number of confusions that occur between the pair.

VI. DISCUSSION

Recognizing functional activities from continuous smartwatch data has profound implications for digital health, behavior monitoring, and medical decision making. Unlike atomic HAR, which identifies basic movements, functional HAR captures richer behavioral patterns that reflect an individual's routines, social engagement, and independence.

Functional activities are central to health measures, including scales used to determine level of independence and need for occupational therapy [57],[58]. Performing such activities is key to the Functional Independence Measure [59], that monitors recovery from stroke and traumatic brain injury. These measures rely on clinician observation or patient self-report, which can be subjective and infrequent. In contrast, this work offers objective, automated tracking of functional activities. This shift has the potential to reduce clinical burden, enable early detection of health changes, manage chronic conditions, and support precision medicine through monitoring and intervention.

This work represents a paradigm shift in HAR by addressing functional activity recognition in the wild, a domain that has remained largely unexplored due to the semantic complexity, inter-individual variability, and real-world sparsity of labeled data. Our key innovations include a novel FT-Transformer-augmented feature representation, which enhances functional HAR by capturing dependencies between engineered features via self-attention. We also introduce the ArWISE dataset, a functional HAR dataset collected from hundreds of participants across multiple study cohorts[3].

We perform a comparison of machine learning models for functional HAR. Models tailored for atomic HAR are not directly included because they are not designed to handle functional activities observed in-the-wild. However, we adapt many of the technologies and include these as baselines for comparison. We demonstrate that our random forest classifier combined with FT-Transformer embeddings outperforms traditional approaches, including classical methods, deep learning methods for i.i.d. and time series data, and methods that utilize various approaches to pretraining. The LOOCV evaluation reveals that performance is impacted by the number of participants, but relative performance between methods remains stable. Furthermore, participant-based splits confirm that the FT+RF model generalizes across individuals without data leakage. Our experimental results also highlight the importance of integrating location and time features into the model to improve functional activity classification.

Despite these advances, there are limitations in the current study. Our LOOCV results highlight that models perform worse when tested on new participants, necessitating the exploration of domain adaptation and minimizing the number of training examples needed through active learning. Our current approach incorporates motion, location, and time. However, additional sensor modalities such as heart rate, temperature, and speech are not utilized, which could further improve recognition. The current approach is also computationally costly. Real-world deployment of the methods will need to address battery constraints, sensor dropout, and hardware heterogeneity.

VII. Conclusion

This work advances the field of human activity recognition (HAR) by shifting the focus from recognizing atomic movements to identifying functional activities. Functional HAR enables the study of behavior patterns critical for health monitoring, intervention design, and population-level analysis. We introduce ArWISE, a set of labeled smartwatch data collected from 503 participants in real-world settings. The methods and dataset introduced in this paper contribute a valuable resource for advancing functional HAR research in diverse, in-the-wild scenarios.

Future research will investigate methods to further fine-tune our models, improving them through additional pretraining, augmentation, and domain adaptation. While this approach currently processes data in the cloud and pushes labels to devices as needed, future versions may streamline the components to fit on edge devices.

Additionally, we will consider leveraging models of atomic movements to generate action-level labels for sensor data. By incorporating these fine-grained action labels into the feature vectors, we aim to improve the representation of functional activities and enhance HAR performance. This approach will provide a richer understanding of human behavior by linking low-level atomic movements with functional activities.

This work represents a significant step toward robust and scalable functional HAR. By demonstrating the effectiveness of a FT-transformer-enhanced random forest model on a large complex dataset, we provide a foundation for future advancements in wearable sensor-based behavior analysis and mobile health.

[3] The dataset, software, and pretrained model are made available to the community at github.com/WSU-CASAS/ArWISE.

Acknowledgment

The authors would like to thank Kayla Ellis for her help with data annotation and thank Maureen Schmitter-Edgecombe, Lisa Wiese, Raven Weaver, Sarah Tomaszewski Farias, and Tricia Winkler for their help in recruiting participants.

References

[1] P. Kumar, S. Chauhan, and L. K. Awasthi, “Human activity recognition (har) using deep learning: Review, methodologies, progress and future research directions,” *Arch. Comput. Methods Eng.*, vol. 31, pp. 179–219, 2024.

[2] H. Haresamudram *et al.*, “Solving the sensor-based activity recognition problem (SOAR): Self-supervised, multi-modal recognition of activities from wearable sensors,” in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 1004–1007.

[3] A. Baldominos, A. Cervantes, Y. Saez, and P. Isasi, “A comparison of machine learning and deep learning techniques for activity recognition using mobile devices,” *Sensors*, vol. 19, no. 3, p. 521, 2019.

[4] G. Somepali *et al.*, “SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training,” *J. Emerg. Technol. Innov. Res.*, vol. 10, no. 7, pp. c859–c863, 2023.

[5] S. Elsayed *et al.*, “Do we really need deep learning models for time series forecasting?,” *arXiv*, vol. arXiv:2101, 2021.

[6] S. Agac and O. D. Incel, “On the use of a convolutional block attention module in deep learning-based human activity recognition with motion sensors,” *Diagnostics*, vol. 13, no. 11, p. 1861, 2023.

[7] D. Thakur *et al.*, “Attention-based multihead deep learning framework for online activity monitoring with smartwatch sensors,” *IEEE Internet Things J.*, vol. 10, no. 20, pp. 17746–17754, 2023.

[8] Y. Wang *et al.*, “A novel deep multifeature extraction framework based on attention mechanism using wearable sensor data for human activity recognition,”

IEEE Sens. J., vol. 23, no. 7, pp. 7188–7198, 2023.

[9] X. Yin, Z. Liu, D. Liu, and X. Ren, "A Novel CNN-based Bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data," *Sci. Rep.*, vol. 12, p. 7878, 2022.

[10] G. Dai, H. Xu, H. Yoon, M. Li, R. Tan, and S.-J. Lee, "ContrastSense: Domain-invariant contrastive learning for in-the-wild wearable sensing," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 4, pp. 1–32, 2024.

[11] Y. Wang, Y. Han, H. Wang, and Z. Xiang, "Contrast everything: A hierarchical contrastive framework for medical time-series," in *International Conference on Neural Information Processing Systems*, 2023, pp. 55694–55717.

[12] N. Ahmad and H. Leung, "ALAE-TAE-CutMix+: Beyond the state-of-the-art for human activity recognition using wearable sensors," in *International Conference on Pervasive Computing and Communication*, 2024, pp. 222–231.

[13] L. Chen, R. Hu, M. Wu, and X. Zhou, "HMGAN: A hierarchical multi-modal generative adversarial network model for wearable human activity recognition," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 3, pp. 1–27, 2023.

[14] W. Huang, L. Zhang, S. Wang, H. Wu, and A. Song, "Deep ensemble learning for human activity recognition using wearable sensors via filter activation," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 15, pp. 1–23, 2022.

[15] X. Zhang *et al.*, "UniMTS: Unified pre-training for motion time series," in *Advances in Neural Information Processing Systems*, 2024, pp. 107469–107493.

[16] S. Miao and L. Chen, "GOAT: A generalized cross-dataset activity recognition framework with natural language supervision," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 4, pp. 1–28, 2024.

[17] L. Arrotta, G. Civitarese, and C. Bettini, "Semantic loss: A new neuro-symbolic approach for context-aware human activity recognition," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 4, pp. 1–29, 2023.

[18] D. N. Jayasimha, S. S. Iyengar, and R. L. Kashyap, "Information integration and synchronization in distributed sensor networks," *IEEE Trans. Syst. Man. Cybern.*, vol. 21, no. 5, pp. 1032–1043, 1991.

[19] S. Ji, X. Zheng, and C. Wu, "HARGPT: Are LLMs zero-shot human activity recognizers?," in *IEEE International Workshop on Foundation Models for Cyber-Physical Systems and Internet of Things*, 2024, pp. 38–43.

[20] A. A. Khaked, N. Oishi, D. Roggen, and P. Lago, "In shift and in variance: Assessing the robustness of HAR deep learning models against variability," *Sensors*, vol. 25, no. 2, p. 430, 2025.

[21] S. Ek, F. Portet, and P. Lalanda, "Transformer-based models to deal with heterogeneous environments in Human Activity Recognition," *Pers. Ubiquitous Comput.*, vol. 27, no. 6, pp. 2267–2280, 2023.

[22] X. Xu *et al.*, "GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 4, pp. 1–34, 2023.

[23] J. Su, Z. Wen, T. Lin, and Y. Guan, "Learning disentangled behaviour patterns for wearable-based human activity recognition," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–19, 2022.

[24] M. Thukral, H. Haresamudram, and T. Ploetz, "Cross-domain HAR: Few-shot transfer learning for human activity recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 1, pp. 1–35, 2025.

[25] A. Mazankiewicz, K. Bohm, and M. Berges, "Incremental real-time personalization in human activity recognition using domain adaptive batch normalization," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–20, 2020.

[26] G. Wilson, J. Doppa, and D. J. Cook, "CALDA: Improving multi-source time series domain adaptation with contrastive adversarial learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14208–14221, 2023.

[27] S. Aminikhanghahi, T. Wang, and D. J. Cook, "Real-Time change point detection with application to smart home time series data," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 1010–1023, 2019.

[28] S. I. Khan *et al.*, "Transition-aware human activity recognition using an ensemble deep learning framework," *Comput. Human Behav.*, vol. 162, p. 183435, 2025.

[29] F. Duan, T. Zhu, J. Wang, L. Chen, H. Ning, and Y. Wan, "A multitask deep learning approach for sensor-based human activity recognition and segmentation," *IEEE Trans. Instrum. Meas.*, vol. 72, p. 2514012, 2023.

[30] H. Qian, S. J. Pan, and C. Miao, "Weakly-supervised sensor-based activity segmentation and recognition via learning from distributions," *Artif. Intell.*, vol. 292, p. 103429, 2021.

[31] S. Xia, L. Chu, L. Pei, J. Yang, W. Yu, and R. C. Qiu, "Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport," *IEEE Trans. Mob. Comput.*, pp. 1–17, 2024.

[32] A. Alazeb *et al.*, "Intelligent localization and deep human activity recognition through IoT devices," *Sensors*, vol. 23, no. 17, p. 7363, 2023.

[33] Y.-H. Kim *et al.*, "MyMove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–21.

[34] W. Ge, G. Mou, E. O. Agu, and K. Lee, "Deep heterogeneous contrastive hyper-graph learning for in-the-wild context-aware human activity recognition,"

Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., vol. 7, no. 4, pp. 1–23, 2024.

[35] Y. Zhang, L. Wang, H. Chen, A. Tian, S. Zhou, and Y. Guo, "IF-ConvTransformer: A framework for human activity recognition using IMU fusion and ConvTransformer," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 88, pp. 1–26, 2022.

[36] S. Salunkhe, S. Shinde, P. Patil, and R. Burchard, "Augmentation approaches to refine wearable human activity recognition," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 603–606.

[37] A. Abedin, M. Ehsanpour, Q. Shi, H. Rezatofighi, and D. C. Ranasinghe, "Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–22, 2021.

[38] S. Shao *et al.*, "ConvBoost: Boosting ConvNets for sensor-based activity recognition," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 2, p. 75, 2023.

[39] N. Ahmad and H. Leung, "ALAE-TAE-CutMix+: Beyond the state-of-the-art for human activity recognition using wearable sensors," in *IEEE International Conference on Pervasive Computing and Communications*, 2023, pp. 222–231.

[40] Z. Hong *et al.*, "CrossHAR: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 1–26, 2024.

[41] N. Oishi, P. Birch, D. Roggen, and P. Lago, "WIMUSim: simulating realistic variabilities in wearable IMUs for human activity recognition," *Front. Comput. Sci.*, vol. 7, p. 1514933, 2025.

[42] Y. Sun, X. Xu, X. Tian, L. Zhou, and Y. Li, "Efficient human activity recognition: A deep convolutional transformer-based contrastive self-supervised approach using wearable sensors," *Eng. Appl. Artif. Intell.*, vol. 135, p. 108705, 2024.

[43] C. I. Tang, I. Perez-Pozuelo, D. Spathis, S. Brage, N. Wareham, and C. Mascolo, "SelfHAR: Improving human activity recognition through self-training with unlabeled data," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–30, 2021.

[44] N. Ahmad and H. Leung, "HyperHAR: Inter-sensing device bilateral correlations and hyper-correlations learning approach for wearable sensing device based human activity recognition," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 1, pp. 1–29, 2024.

[45] O. Napoli *et al.*, "A benchmark for domain adaptation and generalization in smartphone-based human activity recognition," *Sci. Data*, vol. 11, p. 1192, 2024.

[46] A. Reiss, D. Stricker, and G. Hendeby, "Towards robust activity recognition for everyday life: Methods and evaluation," in *Pervasive Computing Technologies for Healthcare*, 2013, pp. 25–32.

[47] S. Chan *et al.*, "CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition," *Nat. Sci. Data*, vol. 11, p. 1135, 2024.

[48] Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Comput.*, vol. 16, no. 4, pp. 62–74, 2017.

[49] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, and J. Danesh, "UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, no. 3, p. 1001779, 2015.

[50] P. M. Butler, J. Yang, R. Brown, M. Hobbs, and A. Becker, "Smartwatch- and smartphone-based remote assessment of brain health and detection of mild cognitive impairment," *Nat. Med.*, 2025, doi: https://doi.org/10.1038/s41591-024-03475-9.

[51] A. A. R. Hassan and M. Feizi-Derakhshi, "ResCBAR-FusionNet: A hybrid CNN-BiGRU-Attention model for human activity recognition," *Signal, Image Video Process.*, vol. 19, p. 675, 2025.

[52] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018, pp. 95–104.

[53] G. Zerveas, S. Jayaram, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," *ACM SIGKDD Conf. Knowl. Discov. Data Min.*, pp. 2114–2124, 2021.

[54] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2d-variation modeling for general time series analysis," in *International Conference on Learning Representations*, 2022.

[55] X. Huang, Z. Khetan, M. Cvitkovic, and Z. Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings," *arXiv:2012.06678*, 2020.

[56] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *International Conference on Neural Information Processing Systems*, 2021, pp. 18932–18943.

[57] R. S. Bucks, D. L. Ashroth, G. K. Wilcock, and K. Siegfried, "Assessment of activities of daily living in dementia: Development of the Bristol activities of daily living scale," *Age Ageing*, vol. 25, no. 2, pp. 113–120, 1996.

[58] C. Graf, "The Lawton instrumental activities of daily living scale," *Am. J. Nurs.*, vol. 108, no. 4, pp. 52–62, 2008.

[59] D. Kidd *et al.*, "The Functional Independence Measure: A comparative validity and reliability study," *Disabil. Rehabil.*, vol. 17, no. 1, pp. 10–14, 1995.