The Physical Activity Assessment Using Wearable Sensors (PAAWS) Dataset: Labeled Laboratory and Free-Living Accelerometer Data

VERONIKA POTTER, Khoury College of Computer Sciences, Northeastern University, USA
HOAN TRAN, Khoury College of Computer Sciences and Bouvé College of Health Sciences, Northeastern University, USA
DANIEL MOBLEY, Brigham and Woman's Hospital, USA
SUZANNE M. BERTISCH, Brigham and Woman's Hospital and Harvard Medical School, USA
DINESH JOHN, Bouvé College of Health Sciences, Northeastern University, USA
STEPHEN INTILLE, Khoury College of Computer Sciences and Bouvé College of Health Sciences, Northeastern University, USA

Poor sleep and sedentary behavior patterns increase the risk of chronic diseases and negatively impact an individual's health and quality of life. Large-scale surveillance studies can unobtrusively measure free-living physical activities, sedentary behaviors, and sleep using wearable sensors; however, many human activity recognition algorithms cannot reliably detect activities in true free-living settings because they are trained on data collected in a controlled, lab setting. We describe the data collection protocol and present the first release of a multimodal, multi-sensor-site dataset (PAAWS R1). The PAAWS R1 release includes ∼4 hours of semi-naturalistic activities from 252 individuals and ∼7 days of 24-hour, free-living activities from 20 adults. We have annotated waking day activities using video to provide second-by-second, ground-truth labels capturing short, quickly changing bouts of activity with realistic activity transitions. Additionally, we have labeled up to two nights of sleep stages from PSG data collected during some nights of the free-living protocol. The PAAWS dataset enables researchers to directly compare activity recognition algorithms on the same participants' data across multiple collection protocols and days of free-living behaviors, encouraging convergence towards robust algorithms that could aid health research and drive novel mobile computing interventions and applications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile devices**.

Additional Key Words and Phrases: Dataset, Physical Activity, Sedentary Behavior, Sleep, Activity Recognition, Accelerometer, Wearables

ACM Reference Format:
Veronika Potter, Hoan Tran, Daniel Mobley, Suzanne M. Bertisch, Dinesh John, and Stephen Intille. 2025. The Physical Activity Assessment Using Wearable Sensors (PAAWS) Dataset: Labeled Laboratory and Free-Living Accelerometer Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 204 (December 2025), 32 pages. https://doi.org/10.1145/3770639

Authors' Contact Information: Veronika Potter, potter.v@northeastern.edu, Khoury College of Computer Sciences, Northeastern University, Boston, USA; Hoan Tran, tran.hoan1@northeastern.edu, Khoury College of Computer Sciences and Bouvé College of Health Sciences, Northeastern University, Boston, USA; Daniel Mobley, dmobley@bwh.harvard.edu, Brigham and Woman's Hospital, Boston, USA; Suzanne M. Bertisch, sbertisch@mgb.org, Brigham and Woman's Hospital and Harvard Medical School, Boston, USA; Dinesh John, dineshjohn@yahoo.com, Bouvé College of Health Sciences, Northeastern University, Boston, USA; Stephen Intille, s.intille@northeastern.edu, Khoury College of Computer Sciences and Bouvé College of Health Sciences, Northeastern University, Boston, USA.

This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2025 Copyright held by the owner/author(s).
ACM 2474-9567/2025/12-ART204
https://doi.org/10.1145/3770639

1 INTRODUCTION

Limited physical activity (PA), excessive sedentary behavior (SB), and/or poor sleep patterns may increase the risks of many chronic diseases (such as heart disease, obesity, and some cancers), and may reduce life expectancy [29, 32]. Valid, scalable methods to estimate PA, SB, and sleep are necessary to advance behavioral health research. Automatic human activity recognition (HAR) systems might make it possible for individuals to passively, and accurately, detect their PA, SB, and sleep via wearable sensors. Developing and validating robust HAR systems to predict true free-living activities from wearable sensor data requires high-fidelity, labeled sensor data. Wearable sensor datasets used for HAR research oftentimes consist of inertial measurement unit (IMU) data (i.e., accelerometer, gyroscope, and magnetometer data) that is acquired passively, continuously, and affordably using research-grade activity monitors, such as ActiGraph accelerometers, or consumer-grade wearable devices, such as smartwatches. Most wearable-based HAR research published to date, however, evaluated models trained using datasets consisting of small sets of relatively simple activities collected in the laboratory (e.g., UCI-HAR [4] and WISDM [30], each of which have only six activities). These algorithms frequently do not perform well on free-living movements [22, 23, 49, 56]. Researchers would benefit from access to more high-quality, free-living, annotated raw-sensor datasets for the purpose of training and validating HAR algorithms.

Although wearable devices are increasingly capable of collecting multimodal data, one of the most important data streams for HAR is accelerometer data. Accelerometers can run continuously with less impact on battery life than other sensors, such as gyroscopes, and these additional sensors have been shown to have only modest improvements in HAR algorithm performance (e.g., [64]). Developing algorithms that perform well on accelerometer-only free-living data, especially from the convenient wrist location, would allow further analysis of data already collected during some of the largest national health surveillance studies. The UK Biobank [24], China Kadoorie Biobank [20], and the National Health and Nutrition Examination Survey (NHANES) [1] collectively include over 130k week-long, wrist-worn-accelerometer-only datasets; these datasets have no ground-truth labels, but they are tied to rich sources of health data and could provide important insights into links between physical activity, sedentary, and sleep behaviors *if the accelerometer data can be analyzed by an accurate, reliable HAR model.*

To convincingly validate such algorithms, the HAR community needs datasets that have second-by-second labeled sensor data of complex, daily life activities (either from a robust lab protocol or true free-living behaviors); such datasets are currently quite limited. Although several labeled, free-living datasets have been collected (e.g., CAPTURE-24 [17], which consists of labeled, free-living wrist accelerometer data labeled at approximately one-minute resolution using time-lapse images for 151 participants collected during a 24 h period), to the best of our knowledge, ours is the first dataset to label at a second-by-second granularity to capture short activity bouts and accurate activity transitions. More public datasets with complex free-living behavior are required to enable researchers to reach consensus on the best-performing HAR algorithms; novel algorithms that have been validated on multiple, independently collected, high-quality labeled datasets can be more confidently used to analyze existing large-scale, unlabeled accelerometer datasets for public health and other purposes.

Over the past four years, we have collected a raw-sensor dataset from 254 participants. A total of 252 participants completed a ∼4 h controlled simulated free-living and laboratory (SimFL+Lab) protocol and 237 participants completed a ∼7 d free-living (FL) protocol during which they may have completed a multi-night sleep protocol collecting polysomnography data (PSG) at home to obtain accurate sleep, wake, and sleep stages labels. We have finished data collection and are in the process of annotating and incrementally releasing data. We anticipate this dataset will complement existing datasets and enable HAR researchers to better directly compare performance of existing algorithms on actual free-living data (with multiple days from single participants). The dataset also contains measurements from sensors on many body positions, including both wrists, enabling new multi-sensor algorithms to be validated on FL data.

In this work, we:

- Provide details on the initial release of the physical activity assessment using wearable sensors dataset (PAAWS R1), an annotated, multimodal-and-sensor-site dataset, that includes SimFL+Lab data from 252 participants (657 h of labeled sensor data) and free-living, daily life activities from 20 adults captured during a week-long FL protocol (∼140 d of FL activity (1,563 h of labeled sensor data) with 29 nights that contain labeled sleep stages and PSG data collected from 15 of the 20 participants).
- Describe the methodology we used to collect, annotate, and quality check PAAWS dataset.
- Present dataset statistics (i.e., number and type of activities and amount of labeled data in the release) and benchmark, using two HAR algorithms, the PAAWS R1 dataset to motivate future work on more robust, generalizable HAR models to improve health interventions and monitoring.

2 BACKGROUND

HAR algorithms for detecting physical activity, sedentary behavior, and sleep are only as useful and generalizable as the data they are trained and evaluated on. Many existing HAR algorithms are trained on datasets comprised of measurements from a small number of participants (i.e., the most popular lab HAR datasets like UCI HAR [4], OPPORTUNITY [21], and DSADS [3] all have data from fewer than 31 people) of a particular demographic (e.g., young adults) in a highly controlled setting (e.g., an exercise laboratory). The activities captured in laboratory settings may not represent the complexity of in-the-wild, free-living movements [41]; and, models trained on such controlled, homogeneous data frequently do not generalize well across different populations and may perform poorly on true free-living data [22, 23, 49, 56]. Additionally, many of the existing HAR models were not compared directly on the same datasets or even validated on multiple HAR datasets, so their performance on more complex populations and activities is unknown. To extend and accurately compare the promising results of lab-driven HAR and ensure generalizability challenges are not encountered when running algorithms on data from free-living settings (such as the datasets in the nation-wide health surveillance studies), the research community needs more publicly available large-scale, raw-sensor datasets with ground-truth labels for PA, SB, and sleep that capture true in-the-wild movements. To address this gap, we have collected a multi-day, multi-sensor, continuously annotated (at a second-by-second granularity) complex laboratory and free-living dataset.

2.1 Laboratory and "Semi-Naturalistic" HAR Datasets

Because supervised physical activity datasets are more cost effective and less laborious to collect than datasets of unsupervised, free-living behaviors, most of the publicly available, labeled, wearable sensor datasets used for HAR were collected in exercise laboratories and/or during supervised protocols. While the method of observation (e.g., video or direct observation) may not directly impact the collected physical activity data [26], recognition results of lab-data-based HAR decreased when evaluated on free-living data [2, 25, 38, 55]. The HAR performance decrease when moving from lab to free-living data can likely be attributed to variations in activities in free-living data (e.g., "walking" in a lab dataset may be comprised only of individuals walking on a flat treadmill with no speed variations whereas "walking" in a free-living dataset may consist of individuals walking quickly to catch a bus or walking while carrying a coffee in their non-dominant hand). In general, lab datasets are composed of a limited, potentially highly constrained, activity set list (summarized in Table 1), providing relatively clean data. Additionally, lab datasets are often comprised of data from a small number of participants from a niche population. Lack of diversity in participants means lack of diversity in data, and the resulting HAR models may not generalize (e.g., HAR models trained on the movements of individuals who were a normal weight did not perform as well on the movements of individuals who were overweight [38]).

To combat some of the limitations that may result from use of existing laboratory-based datasets, we designed the PAAWS SimFL+Lab data collection protocol (described in Section 3.1) to ensure we collected a wide variety

Table 1. Raw accelerometer datasets commonly used in HAR literature collected under supervision in a controlled environment (lab data) or semi-naturalistic setting.

Dataset	*N* =	Demographics	No. Acts	No. Acc Locations	Measurements
ADL [13]	16	11 m, 5 f; 19–83 yo	8	1	A (32 Hz)
Bao + Intille [8]	20	13 m, 7 f; 17–48 yo	20	5	A (76.25 Hz)
CAPPIMU [34]	30	30 m; 20–25 yo	15	9	A (50 Hz), G, M, O
Clemson [39]	30	N/A	3	3	A (15 Hz)
CMU-MMAC [31]	43	N/A	1	5	A (60 Hz), G, AV, O
Cosmo [46]	30	N/A	14	2	A (100 Hz), G, AV, O
DSADS [3]	8	4 m, 4 f; 20–30 yo	19	5	A (25 Hz), G, M
DaLiAc [33]	23	13 m, 10 f; 27 ± 7 yo	13	4	A (204.8 Hz), G
DOMINO [5]	25	17 m, 8 f; 20–59 yo	14	2	A (140-200 Hz), G, M, O
harAGE [36]	19	10 m, 9 f; 41.7 ± 8 yo	11	1	Acc (25 Hz)
KU-HAR [53]	90	75 m, 15 f; 18–24 yo	18	1	A (100 Hz), G
LTMM [28]	71	25 m, 46 w; 65–87 yo	1	1	A (100 Hz)
MHEALTH [7]	10	N/A	12	3	A (50 Hz), HR
MM-DOS [51]	50	50 m; 17–30 yo	4	9	A (50 Hz), G, M, O
MobiAct [63]	57	42 m, 15 f; 20–47 yo	13	1	A (20 Hz), G
MotionSense [35]	24	16 m, 10f; 18–46 yo	6	1	A (50 Hz), G
Nadeem [42]	114	N/A	4	1	A (51), HR
OPPORTUNITY [21]	4	N/A	6	5	A, G, M, O
PAMAP2 [47]	9	8 m, 1 f; 27.22 ±3.31yo	18	3	A (100 Hz), HR
REALDISP [6]	17	N/A	33	7	A (50 Hz)
RealWorld HAR [57]	15	8 m, 7 f; 31.9 ± 12.4 yo	8	7	A (50 Hz)
Swimming [12]	40	25–75 yo	5	1	A, G, M, E
TNDA-HAR [66]	50	25 m, 25 f; 20–35 yo	8	5	A (50 Hz), G, M
UC Berkley MHAD [44]	12	7 m, 5 f; 23–30 yo	11	6	A (30 Hz), O
UC Berkley WARD [67]	20	13 m, 7 f; 19–75 yo	13	5	A (30 Hz), G
UCI HAR [4]	30	19–48 yo	6	1	A (50 Hz), G
UniMiB SHAR [41]	30	6 m, 24 f; 18–60 yo	10	1	A (50 Hz),
USC HAD [69]	14	7 m, 7 f; 21–49 yo	12	1	A (100 Hz)
UT-Data [52]	10	10 m; 23–35 yo	13	2	A (50 Hz), G
UTD-MHAD [19]	8	4 m, 4f	27	1	A, G, M, AV, O
WEAR [11]	22	13 m, 9 f; 27.59 ± 12.4 yo	18	4	A (50 Hz), AV
WISDM [30]	29	N/A	6	1	A (20 Hz)
PAAWS SimFL+Lab	252	137 m, 115 f; 18–80 yo	49	21	A (80 Hz), G, HR, MET

Demographics abbreviations: m – male, f – female, and yo – years old. **Measurement abbreviations**: A – acceleration (acc), AV – audio and/or video, G – gyroscope, HR – heart rate/ECG, M – magnetometer, MET – metabolic breath-by-breath data, O – other sensor measurements (e.g., depth camera, audio, radar, environmental).

of activities common in daily life often not captured in other lab datasets (e.g., many variations on natural walking; a plethora of household chores including vacuuming, sweeping, and cooking; and taking public transit from a transit station). We deliberately included specific variations on common activities (e.g., we included variations of walking including indoor walking such as *Treadmill_3mph_Drink_Lab* and walking outside such

as walking while on the phone, unrestricted, and on a variety of surfaces). Further, the PAAWS SimFL+Lab dataset is comprised of data from 252 individuals ages 18-80, roughly split by gender, capturing a wide diversity of movements that may, hopefully, allow HAR models to generalize to many able-bodied individuals from a variety of demographics.

2.2 Free-living HAR Datasets

Because the data collection process for free-living (FL) data is expensive, laborious, and time intensive in comparison to collecting supervised, laboratory data, there are fewer publicly available, raw-sensor, FL HAR datasets (summarized in Table 2).

Table 2. Summary of raw accelerometer datasets collected in free-living environments.

Dataset	**Labels?**	***N* =**	**Length (d)**	**Sensor Locations**	**Measures**	**Sleep Labels?**
In-hospital PA [50]	No	58	2 – 10	Wrist	A, G, E	N/A
LTMM [28]	No	71	3	Lower back	A	N/A
NHANES [1]	No	∼16k	7	Wrist	A	SR (‘17-‘18)
China Kadoorie Biobank [20]	No	∼20k	7	Wrist	A	SR*
UK Biobank [20]	No	∼100k	7	Wrist	A	SR*
Leisure Activities [9]	Partially	6	5	Wrist	A, E	N/A
UbiComp ‘08 [27]	Partially	1	16	Wrist, pocket	A	N/A
CAPTURE-24 [17]	Yes	151	1	Wrist	A	SR
ExtraSensory [62]	Yes†	60	2 – 28	Wrist, phone	A, G, E, O	SR
MMASH [48]	Yes†	22	1	Wrist	A, O	SR
PAAWS R1 FL	Yes	20	7	Wrists, waist, thigh, ankle, phone	A	SR (all nights), expert labeled sleep stages from PSG (up to two nights)

† Labels from self-report diaries

* Quality and length, not wake/sleep times.

Abbreviations: A – acceleration, G – gyroscope, E – ambient environmental measurements (e.g., temp, light), O – other sensor measurements, SR – self-report wake/sleep times (from diaries, EMAs, or recall with a researcher).

Most of these datasets contain data from only one sensor (frequently a wrist-worn accelerometer) likely chosen for high wear compliance in comparison to other sensor wear locations [40, 61]. The labeled, true free-living datasets include CAPTURE-24 [17], ExtraSensory [62], and MMASH [48]. The latter two datasets' labels come from participants' self-reporting their behaviors and are at a temporal resolution of the participants' choice. To remedy some of the bias and issues that may exist in self-report labels, such as inaccurate timestamps, the researchers who collected CAPTURE-24 labeled the data using time-lapsed still images taken every 20-40 s; activities in CAPTURE-24 are annotated with, at best, one minute resolution. Additionally, reliance on time-lapse images may result in situations where there is insufficient context to label activities or the activity's precise start/end times [17]. While an improvement from self-report labels, the granularity used in CAPTURE-24 may not be sufficient to capture precise activity transitions or short activities. Further, the CAPTURE-24 dataset only includes data from 24 continuous hours per participant, preventing researchers from investigating behavioral trends over the course of the week, differences in work-week days versus weekend days, or differences due to

small changes in how sensors were placed on the body over different days. In general, these publicly available, labeled, free-living datasets have moved the community towards generalizable FL HAR. Results from validating HAR algorithms trained on laboratory data, however, suggest that acceleration signals from a single-wrist sensor may not capture the complexity of many movements [43, 45], thus researchers may need to explore other wear locations and combinations of sensor wear locations in FL HAR. To improve the performance of HAR algorithms on FL data and determine the best sensor wear locations for long-term, reliable HAR, researchers require large, representative (i.e., data collected from weekdays and weekends, and from people of different ages), and continuously labeled multi-sensor-location FL HAR datasets. Because the PAAWS FL data is a labeled, week-long, multi-sensor-site free-living dataset containing data with demographic diversity, we hope this dataset will allow researchers to investigate some of these questions.

3 DATA COLLECTION AND ANNOTATION METHODOLOGY

The PAAWS dataset consists of annotated, raw-sensor data collected from the same participants across two protocols: the simulated free-living and exercise laboratory (SimFL+Lab) protocol and the free-living (FL) protocol.

3.1 Simulated Free-Living and Exercise Laboratory Protocol

The simulated free-living and exercise laboratory (SimFL+Lab) protocol was a ∼4 h supervised protocol consisting of a ∼1 h (minimally controlled) simulated free-living (SimFL) protocol with 16 "activities" and a ∼3 h (highly controlled) exercise laboratory (Lab) protocol with 33 activities. During the SimFL protocol, participants engaged in a sequence of events designed to approximate true free-living behavior, and participants were asked to complete the activities as they would in their daily life. Activities took place on Northeastern University's Boston campus and in the surrounding city, which resulted in many bouts of realistic walking and daily life multi-tasking activities, such as scrolling through a phone while taking public transportation or having a conversation while walking. The Lab protocol was a highly controlled protocol derived from previous work [58] with modifications. Every participant was given specific instructions for how to complete common exercise laboratory activities ranging from sedentary behaviors (e.g., sitting still, lying on back), to light/moderate activity (e.g., walking on a treadmill or biking), to vigorous activities (e.g., running on a treadmill). We also included activity variations that have not been included in prior physical activity datasets but are relevant to daily life (e.g., walking on a treadmill with a drink in hand, and standing while conversing or chopping food).

During the SimFL+Lab protocol, we collected acceleration, inertial measurement unit (IMU), heart rate, and energy expenditure data. Throughout the entirety of the protocol, participants wore 20 ActiGraph GT9X Link sensors (ActiGraph, LLC), which collected both accelerometer and IMU data, and a Polar H7 chest strap heart rate monitor (Polar Electro, Inc.), which collected heart rate data. The primary ActiGraph accelerometers sampled at a rate of 80 Hz (range ±8 g). The sensors collecting IMU data sampled at a rate of 100 Hz and included measurements of a secondary accelerometer (range ±16 g), gyroscope (range ±2000 degree/s), and magnetometer (range ±4000 μT). On both the left and right side of the body, the ActiGraph sensors were placed as follows: four sensors around each ankle (one in each of the lateral, anterior, posterior, and medial positions); one sensor in the middle of each thigh; three sensors on each side of the waist (one on the anterior axillary line, one on the midaxillary line, and one on the posterior axillary line); and two sensors on each wrist (one on the top, like a wristwatch is commonly worn, and one on the bottom). Sensor placement for the SimFL+Lab protocol is shown in Fig. 1. Accelerometer and IMU data were also collected from an additional ActiGraph GT9X Link sensor that was attached to the upper back of the participant's phone; accelerometer and IMU data were collected from a total of 21 ActiGraph sensors. During 13 of the 33 activities completed in the Lab portion of the protocol, participants

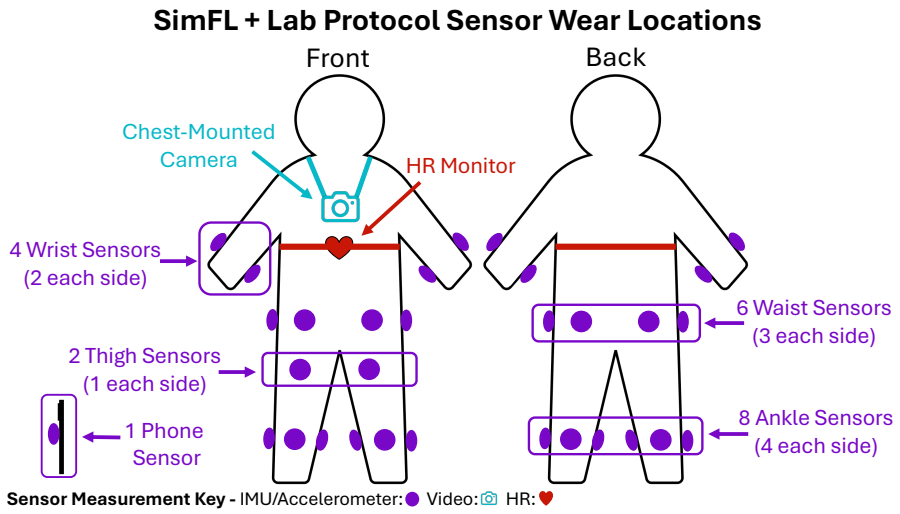

Fig. 1. An overview of the sensor wear locations and data collected during the SimFL+Lab protocol.

wore a tight-fitting mask attached to a COSMED Quark CPET metabolic cart (COSMED srl) to measure breath-by-breath energy expenditure. We computed the metabolic equivalent of task (MET) from breath-by-breath energy expenditure for these activities.

Because sensor location impacts the performance of HAR models [58], we designed our SimFL+Lab protocol to take measurements from the main sensor-wear locations used in HAR that could generalize easily to a free-living setting: wrists, hip/waist, thigh, and ankle. Additionally, because small displacements can impact recognition rates, we considered a variety of sensor placements at each wear-location (shown in Fig. 1). This sensor arrangement was feasible only for the SimFL+Lab protocol because of the short protocol duration and because a research assistant was available throughout the entire protocol to verify that sensors were correctly placed (initially) and remained in place throughout the protocol (i.e., a research assistant could check if sensors fell off and needed to be reapplied). Further, because modality affects HAR results [18] and the SimFL+Lab protocol lasted for a short amount of time, we collected additional IMU data from each ActiGraph sensor. Not all these sensors could be used in the long-term FL setting data collection. The IMU sensor drains the battery of the device rapidly, making it unsuitable for long-term wear. The chest-worn Polar HR device becomes uncomfortable after long wear periods and dries out preventing the device from functioning. And, the COSMED device used to measure energy expenditure via indirect calorimetry cannot be used in a non-lab environment because the equipment is cumbersome. Wearing multiple sensors around the wrist is viable for a short time but becomes uncomfortable in daily free-living life. The SimFL+Lab sensor suite was manageable for several hours but not for prolonged wear, thus we used a subset of only six of the SimFL+Lab sensor wear locations during the FL protocol.

3.2 Free-Living Protocol

The FL protocol was a ~7 d, unsupervised, uncontrolled free-living period. Participants went about their lives as usual while wearing accelerometers on multiple limbs. Participants also wore a front-facing, chest-mounted Cammpro 3 camera attached to an extended battery. The camera hung around the neck, partially stabilized by a magnet worn under the participant's shirt on a second lanyard (i.e., the camera was held in place by a magnet under the participant's shirt). Representative thumbnails of the type of video footage acquired are shown in Fig. 3.

This protocol resulted in multi-limb accelerometer data from a week of realistic, habitual, free-living behaviors from both weekdays and weekends, including sleep and up to two nights of accurately labeled sleep and sleep stages.

During the FL protocol, participants wore five ActiGraph GT9X Link sensors placed on the tops of the left and right wrists (like a watch), on the right waist following the anterior-axillary line, in the middle of the right thigh, and laterally just above the right ankle (Fig. 2). To ensure the ActiGraph devices would function for the entire FL protocol without requiring charging, we set them to record accelerometer data using the primary accelerometer (sampling at 80 Hz with a range of ±8 g), but not full IMU data (as in the SimFL+Lab protocol) because IMU sensing would drain the battery within a day. All five on-body accelerometers were worn throughout the protocol except during water-based activities (e.g., bathing or swimming), when participants removed the wrist, ankle, and hip sensors; the thigh sensor remained on because it was attached to the skin of the thigh using waterproof medical tape. An additional sixth ActiGraph GT9X Link sensor was attached to the back of the participant's phone. These six locations were chosen because they are commonly used in prior work developing machine learning models [54] and, with the exception of the thigh sensor, are practical for long-term monitoring [40]. Additionally, these six wear locations directly correspond to locations of sensors worn during the SimFL+Lab protocol, allowing for the development of transfer learning approaches that may train an algorithm on one dataset and evaluate on another. We did not, however, have participants wear the entire sensor suite used in the SimFL+Lab protocol during the FL protocol because it would impose significantly more discomfort, and we were concerned about the battery life of the devices. In addition to the body-worn, research-grade sensors, we collected phone data (i.e., application usage and detected steps) via an Android application from participants who owned Android devices. On each day during the FL protocol, participants met virtually with a research assistant to report their sleep quality, time-in/out-of-bed, accelerometer non-wear, and any activities when participants did not wear the front-facing camera (e.g., during showering or bathing). During this time, we ensured all equipment was functioning as intended and worn correctly by participants. The dataset includes notes about situations that might impact data quality.

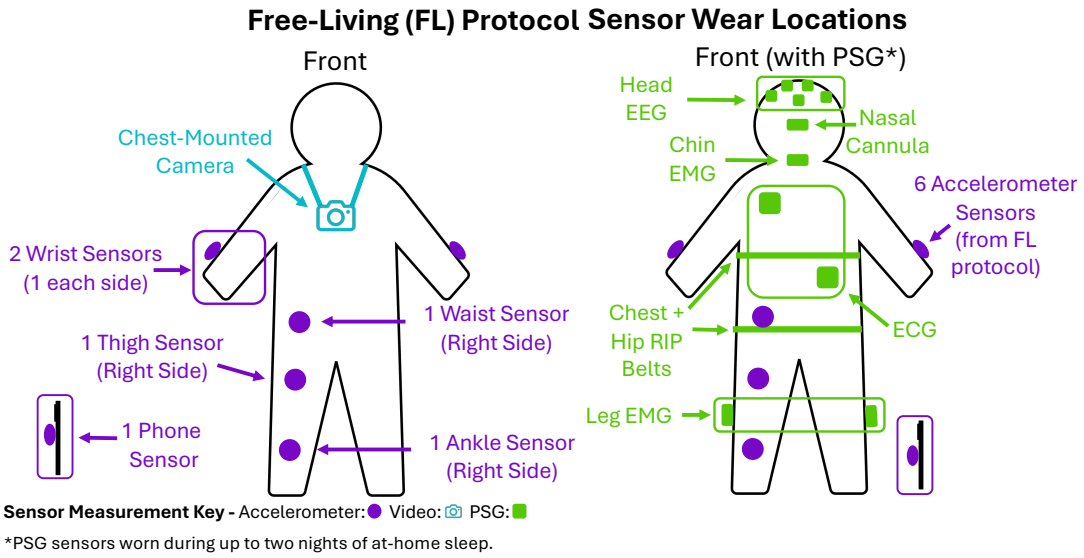

Fig. 2. Sensor wear locations and data collected during the FL protocol at home during free-living, in-the-wild behaviors.

3.2.1 Obtaining Accurate Sleep Labels During the Free-Living Protocol. For up to two nights during the FL protocol, PSG data were collected while the participant slept, at home, concurrently with the sensors described in Section 3.2. During each night the participants collected PSG data, they slept while connected to a Nox A1 portable PSG system (Nox Medical, Inc.) with the following sensors: left and right leg electromyography (EMG), ten-lead electroencephalogram, pulse oximeter, and two-lead electrocardiogram. The PSG data were then annotated according to the protocols outlined in Section 3.4 to provide accurate sleep, wake, and sleep stage annotations. In addition to the labels, we release the raw PSG signals.

3.3 Participants

The PAAWS R1 dataset contains continuous raw sensor data from 252 participants (20 in FL) with second-by-second annotation. Of these participants, 116 were ages 18-34 (12 in FL), 62 were ages 35-54 (one in FL), 74 were ages 55+ (seven in FL), and 115 identified as female (nine in FL). Participants in the PAAWS study were recruited using physical flyers, online social media platforms, and word of mouth. Each participant in the PAAWS study consented to participation by reading through and signing a consent form with a researcher. For their efforts, each received up to $750 (USD). Compensation was determined based on the protocols each participant completed. The PAAWS study protocol was approved by the Institutional Review Board at Northeastern University.

3.4 Annotation Protocols: Labeling the SimFL+Lab, FL, and Sleep Data

The PAAWS SimFL+Lab and FL data were annotated at a second-by-second granularity from video footage captured during the two protocols. Although laborious, hand labeling video footage is the gold standard for establishing ground truth behavioral labels. Each annotation for the PAAWS dataset consists of at least two labels: the participant's posture and PA type. Additionally, to help researchers (who will not have access to the video data) interpret results, annotations may include non-mutually exclusive labels that provide additional context about the PA type: the high-level behavior (HLB) and/or contextual parameters (CP). The annotation scheme was developed based on discussions with members of the Sensor Methods Colaboratory and published literature [61].

The SimFL+Lab and FL datasets were annotated by trained undergraduate students from Northeastern University. Each annotator completed a ~6 h training protocol. Before beginning the annotation training, each annotator completed the required university training for handling human subject data. Before annotating new data, annotators were required to achieve at least a 90% second-by-second agreement for PA type and posture labels and 80% second-by-second agreement for HLB and CP labels with a gold-standard label set across short training videos. These training videos include free-living footage capturing common daily living PAs. We calculated inter-rater reliability among annotators post-training using additional clips collected from both the SimFL+Lab and the FL protocols. The SimFL+Lab clips represent a standard data collection session. The FL clips include simple and obvious activities with few transitions occurring, but also situations where a person is performing complex PAs and transitioning quickly between them. We computed Cohen's Kappa for the mutually exclusive labels (posture and PA type) and Krippendorff's Alpha for the non-mutually exclusive labels (CP and HLB) across our annotators for both the SimFL+Lab and FL data. For the SimFL+Lab scenario, the Kappa value of our PA type and posture labels are $\kappa = 0.89$ and $\kappa = 0.79$ respectively (strong and moderate-to-strong agreement). For the simpler FL scenario, we computed a Kappa value of $\kappa = 0.76$ and $\kappa = 0.79$ for PA type and posture labels (moderate-to-strong agreement) and an Alpha value of $\alpha = 0.61$ and $\alpha = 0.39$ for the HLB and CP labels (moderate and low-to-moderate agreement). For the more complex FL scenario, we computed a Kappa value of $\kappa = 0.61$ and $\kappa = 0.57$ for PA type and posture labels (moderate agreement) and an Alpha value of $\alpha = 0.67$ and $\alpha = 0.40$ for the HLB and CP labels (moderate and low-to-moderate agreement). Because annotators were trained to label only what they were sure of, variability may have resulted from the confidence of annotators when identifying different activities, postures, or contextual parameters. Additionally, to standardize many aspects of the labels,

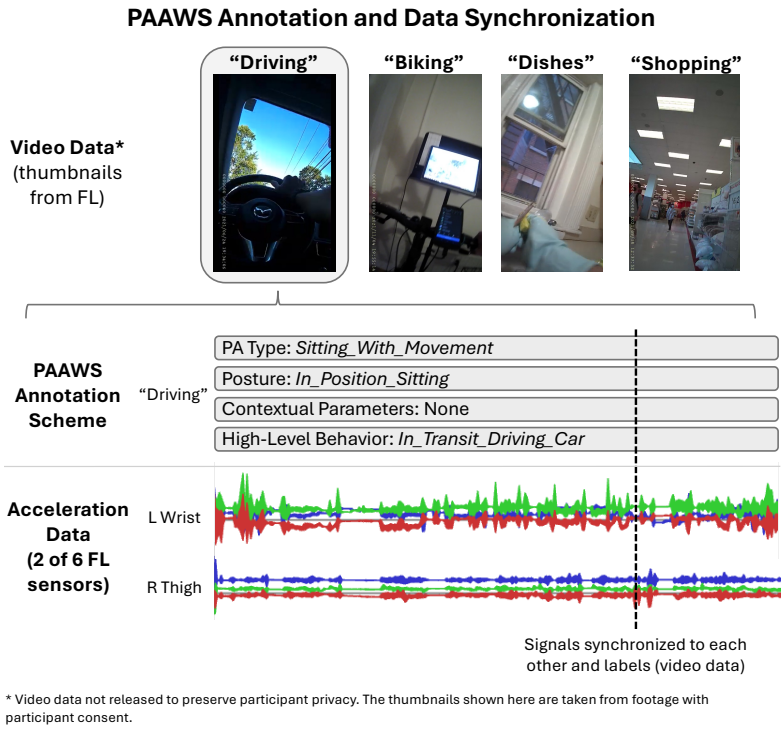

Fig. 3. Video-based retrospective annotation of the first-person footage using the PAAWS annotation scheme. Data streams are synchronized to each other and the footage. **Abbreviations**: L – left and R – right.

immediately after annotating the actual PAAWS data, annotators addressed potential quality errors flagged by custom software, and all annotations went through a final cleaning and quality check phase before being released (detailed in our codebook at github.com/mHealth-Research-Group/paaws-benchmarking). Because of the sensitivity of labeling video footage that cannot be deidentified, annotators worked in a room only accessible by individuals who were involved in the project and completed special training on handling the potentially sensitive data. To further promote participant privacy, each annotator only viewed the footage they were directly responsible for annotating. We summarize the annotation process and scheme in Fig. 3.

All sleep data were annotated by registered polysomnographic technologists at Brigham and Women's Hospital following protocols outlined by the American Academy of Sleep Medicine [10]. Sleep stages were labeled at 30 s temporal granularity with other associated events labeled with sub-second precision. Although the non-PSG nights do not include 30 s ground truth sleep state labels, the dataset includes self-report sleep quality, time-in-bed (start and end), and time asleep (start and end) for every night in the FL protocol.

3.5 Data Processing: Synchronizing Measurements and Labels

We synchronized the data collected from a participant to correct for sensor internal clock drift after annotation for that participant was complete. We corrected time drift between sensors using synchronization points in the accelerometer data; synchronization points were marked by performing a "box drop" (i.e., dropping a box containing all the sensors from a small height, resulting in a distinct spike across all sensors). To correct for time

drift between camera footage, annotations, and sensor data, participants performed a synchronization activity at the start and end of each day in the FL protocol by clapping their hands and hugging their knees five times in camera view. Finally, we synchronized the PSG data with the other sensors using posture changes throughout the night indicated by acceleration shifts in data collected from the chest-mounted PSG accelerometer and limb-worn accelerometers in the FL protocol. Before release, synchronized data were reviewed by a researcher (who did not perform the initial syncing) to verify sensor temporal alignment was achieved during the synchronization procedure.

4 VALIDATION: DATASET STATISTICS AND BENCHMARKING

To characterize the PAAWS R1 SimFL+Lab and FL datasets, we computed overall dataset statistics (summarized in Table 3 and Table 4) and benchmarked the dataset with a random forest model (RF), a previously well-performing HAR model (Table 5, 6, and 7).

4.1 PAAWS R1 (SimFL+Lab and FL) Dataset Statistics

The PAAWS SimFL+Lab dataset contains 657.3 h of annotated[1] raw-sensor data acquired from 252 different participants (155.8 ± 18.4 min of labeled data from each participant). The dataset also contains phone data from 28 participants' personal Android devices (three participants who collected Android phone data during the SimFL+Lab protocol have some data loss). In addition to the 33 exercise laboratory activities that each participant completed after being given an identical set of instructions, this dataset includes 22 other physical activity labels capturing either simulated free-living or transitionary movements. As such, the PAAWS R1 SimFL+Lab dataset is one of the most complex and diverse semi-controlled raw sensor HAR datasets available.

Table 3. A summary of the annotations for the SimFL+Lab and FL protocols in the PAAWS R1 dataset. We define an activity as a unique combination of a distinct PA type label and possible posture, CP, and/or HLB label.

					Distinct Label			
Protocol	**Data (h)**	**PA Labels (h (%))**	**Bouts of Activity**	**Unique Activities**	**PA Type**	**Posture**	**CP**	**HLB**
SimFL+Lab	717.89	657.29 (91.5%)	34,211	1,025	55*	8	277	–
FL	3,288.32	1,563.61 (47.6%)	45,130	5,241	38	8	473	378

* 33 activity labels in the SimFL+Lab annotations are unique to the laboratory protocol and only used in this dataset.

The PAAWS R1 FL dataset contains 1,563.6 h of continuously annotated, second-by-second raw sensor data from eight days and seven nights of 20 participants' daily movements (78.18 ± 11.49 h of labeled data per participant, ranging from 54.18 h (33.5%) to 102.55 h (59.0%) of labeled sensor data per participant) and data from two participants' personal Android devices. The PAAWS R1 FL dataset captures more than five thousand different variations on activities—a unique combination of PA type, posture, CP, and HLB labels; we summarize some variations in Table 4. The variability in the number of activities and length of the FL protocol can be attributed to differences in participant lifestyles. The dataset contains thirty-eight different physical activity type labels and eight postures; the contexts in which participants conducted many of these activities vary substantially. Overall, 45,130 different bouts of activities are labeled. The minimum duration of an activity bout is 1 s, with 50% of all bouts lasting less than 34 s, and 75% of activity bouts lasting less than 90 s. The length of these activity bouts underscores the importance of labeling free-living data at a second-by-second granularity to capture the quick, rapidly changing activities individuals do daily.

[1]We define annotated data as data that include a distinct, usable PA type label (i.e., the annotation contains a PA type label that is not *Too_Complex*, *Video_Unavailable/Indecipherable*, or *Other*).

Table 4. Notable behaviors encompassing greater than one hour of labeled data in the PAAWS R1 FL dataset. Above the fold: eight of the twenty most common activities labeled in the dataset highlighting the diversity variations in activities. Below the fold: six of the one hundred least common activities. Duration is reported in minutes.

Activity	**Total Duration**	**No. of Participants**	**Mean (SD) Duration***	**Maximum Duration***	**Minimum Duration***
Sitting, on Computer	4,005.89	18	269.72 (209.58)	682.22	1.00
Sitting, on Phone	2,513.96	20	125.77 (146.69)	573.23	0.72
Sitting, Watching TV	1,627.42	20	82.87 (73.47)	323.85	2.25
Driving	1,500.04	7	222.60 (198.34)	562.47	44.08
Walking	1,188.08	20	60.36 (46.68)	173.71	0.48
Lying, on Phone	839.16	15	56.22 (69.00)	209.25	0.50
Walking, w Bag	695.73	20	37.00 (38.11)	138.85	1.83
Cooking	646.57	19	47.62 (38.26)	129.22	0.28
Lying, TV + Phone	97	4	24.25 (17.54)	38.68	1.97
Standing, Talking	94.34	9	10.48 (13.46)	39.88	0.45
Drinking, Sitting	85.33	15	5.73 (11.71)	43.4	0.1
Chores	75.71	15	5.39 (8.35)	27.04	0.27
Dressing/Undressing	68.27	17	4.03 (4.59)	19.95	0.62
Standing, Grooming	61.84	16	3.87 (8.96)	37.18	0.12

* Mean, maximum, and minimum minutes of activity are computed only from the participants that completed that activity.

Activity abbreviations (defined by the PAAWS annotation scheme): ***"Sitting, on Computer"*** (PA Type: *Sitting_With_Movement*; Posture: *In_Position_Sitting*; CP: *Browsing_Computer*, *Located_Indoors*; HLB: *Using_Computer/Screen*), ***"Sitting, on Phone"*** (PA Type: *Sitting_With_Movement*; Posture: *In_Position_Sitting*; CP: *Texting/Browsing_Phone*, *Located_Indoors*), ***"Sitting, Watching TV"*** (PA Type: *Sitting_With_Movement*; Posture: *In_Position_Sitting*; CP: *Located_Indoors*; HLB: *Watching_TV/Movies*), ***"Driving"*** (PA Type: *Sitting_With_Movement*; Posture: *In_Position_Sitting*; CP: *Located_Outdoors*; HLB: *In_Transit_Driving_Car*), ***"Walking"*** (PA Type: *Walking*; Posture: *In_Position_Upright*; CP: *Located_Outdoors*; HLB: *In_Transit_Active*), ***"Lying, on Phone"*** (PA Type: *Lying_With_Movement*; Posture: *Lying_On_Back*; CP: *Texting/Browsing_Phone*, *Located_Indoors*; HLB: *Relaxing_Passive*), ***"Walking, w Bag"*** (PA Type: *Walking*; Posture: *In_Position_Upright*; CP: *Located_Outdoors*, *Carrying_Bag_On_Shoulder/Back_Light*; HLB: *In_Transit_Active*), ***"Cooking"*** (PA Type: *Puttering_Around*; Posture: *In_Position_Upright*; CP: *Located_Indoors*; HLB: *Cooking/Prepping_Food*), ***"Lying, TV + Phone"*** (PA Type: *Lying_With_Movement*; Posture: *Lying_On_Back*; CP: *Texting/Browsing_Phone*, *Located_Indoors*; HLB: *Relaxing_Passive*, *Watching_TV/Movies*), ***"Standing, Talking"*** (PA Type: *Standing_With_Movement*; Posture: *In_Position_Upright*; CP: *Located_Outdoors*, *In_Conversation*; HLB: *Meeting_Social_Gathering_In_Person*), ***"Drinking, Sitting"*** (PA Type: *Sitting_With_Movement*; Posture: *In_Position_Sitting*; CP: *Holding_Drink*, *Located_Indoors*; HLB: *Eating/Dining*), ***"Chores"*** (PA Type: *Puttering_Around*; Posture: *In_Position_Upright*; CP: *Holding_Other_Object_Light*, *Located_Indoors*; HLB: *Doing_Common_Housework_Light*), ***"Dressing/Undressing"*** (PA Type: *Standing_With_Movement*; Posture: *In_Position_Upright*; CP: *Located_Indoors*, *Holding_Other_Object_Light*; HLB: *Dressing/Undressing*), ***"Standing, Grooming"*** (PA Type: *Standing_With_Movement*; Posture: *In_Position_Upright*; CP: *Located_Indoors*; HLB: *Grooming/Self_Care*).

Although the PAAWS annotation protocol allows second-by-second annotation of activities, there are moments in the chest-mounted camera video footage when the activity is not distinguishable. When annotators are unable to discern the activity, they use *Too_Complex*, or *Video_Unavailable/Indecipherable*, or *Other* labels for PA types, postures, CPs, and HLBs. We included these labels in our taxonomy to allow our annotators to capture the full breadth of free-living movements without compromising the participants' behaviors. Free-living activity encompasses many activities that, for a variety of reasons such as time or privacy, are not captured in controlled PA datasets (e.g., using the restroom, entering private information on a computer screen, and attending a meeting

where individuals do not want to be recorded). To ensure our participants were able to conduct their daily lives with comfort, participants were instructed to cover or remove the camera during any moments they had privacy concerns. To prevent an ambiguous gap in the annotation, we needed to ensure we could document these instances and chose to do so with the *Video_Unavailable/Indecipherable* label. Additionally, although we developed our annotation taxonomy to cover most free-living activities, there are (rare) instances that participants might be doing an activity the taxonomy could not adequately describe. Again, to ensure these did not result in an ambiguous gap in annotation (e.g., is the annotation missing or is there nothing to annotate?), we created the *Other* and *Too_Complex* labels. The *PA_Type_Other* label is used when a participant was doing a recognizable activity not included in our taxonomy and the *PA_Type_Too_Complex* label is used when a participant was doing a rapidly changing activity that the annotator did not believe they could accurately annotate. We excluded these labels from our computations of "labeled data." The *PA_Type_Video_Unavailable/Indecipherable* accounts for 326.3 h (9.9%) of all data collected, and the *PA_Type_Other* and *PA_Type_Too_Complex* labels combined account for 7.3 h (2.2%) of all raw data collected in the FL protocol. In the instances when annotators cannot decipher the PA type or posture because the participants removed the camera for a specific activity, such as *Showering* or *Bathing*, these activities were reported each day during the daily check-in and activity recall with a research assistant; when these activities were reported, we included these notes/annotations in the dataset.

4.2 Benchmarking the PAAWS R1 Dataset

We benchmark a HAR algorithm on the PAAWS R1 dataset for both waking day activities (SimFL+Lab and FL datasets) and sleep (FL dataset only). For both the SimFL+Lab and FL benchmarking, we report the results of a random forest (RF) algorithm because RFs have been used frequently used in other research with good results [16, 17, 65]. When benchmarking the waking day activities, we considered two sensor-wear locations—the left wrist (top) and right thigh. For the waking day HAR experiments, we used the same preprocessing and evaluation protocols across the SimFL+Lab and FL datasets. We considered only the accelerometer data sampled at 80 Hz and extracted 59 features from segmented 10 s, non-overlapping windows (Appendix Table 9). We evaluated each RF using leave-one-participant-out cross-validation (LOPO CV) and reported the weighted and macro F1 scores to account for the class imbalance within the data. We summarize the results of our experiments in Tables 5, 6, and 7. More details about features, model parameters, and class-wise model performance can be found in the Appendix. We benchmark the sleep labels collected during the FL protocol using the existing SWaN algorithm [59] on both wrist sensors and report the results of this experiment in Table 8. The source code for reproducing our results, and the model weights from our experiments, are available: github.com/mHealth-Research-Group/paaws-benchmarking.

4.2.1 Benchmarking the SimFL+Lab Dataset. To benchmark a HAR algorithm using the SimFL+Lab dataset, we consider three activity recognition tasks: (1) using a re-grouped activity set by pre-defining mappings to simplify the raw labels into a traditional HAR problem of five activities like those found in other lab datasets, (2) using a re-grouped activity set of nine activities, and (3) utilizing the raw labels (i.e., 42 total activities[2]), to capture the diversity of activities present in the dataset. The first task, using re-grouped activity labels, is "easier" than the

[2] *Ab_Crunches_Lab*, *Arm_Curls_Lab*, *Chopping_Food_Lab*, *Cycling_Active_Pedaling_Regular_Bicycle*, *Folding_Clothes*, *Lying_On_Back_Lab*, *Lying_On_Left_Side_Lab*, *Lying_On_Right_Side_Lab*, *Lying_On_Stomach_Lab*, *Machine_Chest_Press_Lab*, *Machine_Leg_Press_Lab*, *Organizing_Shelf/Cabinet*, *Playing_Frisbee*, *Push_Up_Lab*, *Push_Up_Modified_Lab*, *Sit_Recline_Talk_Lab*, *Sit_Recline_Web_Browse_Lab*, *Sit_Typing_Lab*, *Sit_Writing_Lab*, *Sitting_Still*, *Sitting_With_Movement*, *Stand_Conversation_Lab*, *Stand_Shelf_Load_Lab*, *Stand_Shelf_Unload_Lab*, *Standing_Still*, *Standing_With_Movement*, *Stationary_Biking_300_Lab*, *Sweeping*, *Treadmill_2mph_Lab*, *Treadmill_3mph_Briefcase_Lab*, *Treadmill_3mph_Conversation_Lab*, *Treadmill_3mph_Drink_Lab*, *Treadmill_3mph_Free_Walk_Lab*, *Treadmill_3mph_Hands_Pockets_Lab*, *Treadmill_3mph_Phone_Lab*, *Treadmill_4mph_Lab*, *Treadmill_5.5mph_Lab*, *Vacuuming*, *Walking*, *Walking_Down_Stairs*, *Walking_Up_Stairs*, and *Washing_Dishes_Lab*.

second. For the first grouped activities task, we considered five activities: *sitting*, *standing*, *lying down*, *walking*, and *biking*. To add complexity to the task, we added four activities that capture slightly more of the breadth of daily life activities: *walking upstairs*, *walking downstairs*, *gym exercises*, and *household chores*. For the unmapped task, we considered 42 different PA types. We validate the results of a HAR model on each task across three different participant set sizes: 20 participants (that have corresponding FL data), 126 participants (50% of the dataset)[3], and 244 participants (the entirety of the dataset after removing participants with known sensor issues or missing data). To demonstrate how the SimFL+Lab dataset can be used in conjunction with the FL dataset, we evaluated the RF models trained on the SimFL+Lab data on the FL data for the recognition tasks using the re-grouped activity classes containing five and nine activities (Table 6). Confusion matrices for each recognition task are reported in the Appendix (Fig. 5, 6, and 7).

Table 5. Benchmarking results by number of participants and sensor location for the RFs trained and evaluated using the SimFL+Lab data for regrouped and raw activities (5, 9, and 42 activities). Metrics are reported as the mean ± standard deviation for leave-one-participant-out cross validation.

Sensor	***N* =**	**5 Activities (Regrouped)**		**9 Activities (Regrouped)**		**42 Activities (Raw)**	
		W F1	**M F1**	**W F1**	**M F1**	**W F1**	**M F1**
Left Wrist (Top)	20	0.75 ± 0.05	0.69 ± 0.08	0.69 ± 0.06	0.54 ± 0.08	0.41 ± 0.04	0.28 ± 0.07
	126	0.80 ± 0.05	0.74 ± 0.07	0.74 ± 0.05	0.60 ± 0.09	0.47 ± 0.07	0.35 ± 0.08
	244	0.80 ± 0.05	0.75 ± 0.07	0.75 ± 0.06	0.62 ± 0.09	0.48 ± 0.06	0.37 ± 0.08
Right Thigh	20	0.90 ± 0.02	0.85 ± 0.04	0.84 ± 0.03	0.75 ± 0.05	0.45 ± 0.05	0.29 ± 0.05
	126	0.90 ± 0.08	0.85 ± 0.08	0.85 ± 0.08	0.78 ± 0.10	0.47 ± 0.07	0.33 ± 0.07
	244	0.91 ± 0.06	0.86 ± 0.07	0.86 ± 0.06	0.79 ± 0.09	0.48 ± 0.07	0.35 ± 0.07

Abbreviations: W F1 – weighted F1 score and M F1 – macro F1 score.

In line with prior work, the RFs trained on the right thigh, a lower body sensor, perform better than the models trained and evaluated on wrist data when evaluated (e.g., [37]) on the activity sets comprised of five and nine activities. Considering the confusion matrices (Appendix Fig. 5 and 6),this can likely be attributed to the thigh sensor data better capturing posture and lower body movements that differentiate many of the movements in these classification tasks. Additionally, across both sensors, the performance of the RF decreases as activities are added, indicating an inability of the RF or the feature set to generalize to more complex HAR tasks. The performance of the RFs was particularly poor when we consider the raw PA labels (i.e., 42 total activities) that include slight variations of similar activities (e.g., walking at different speeds while carrying objects and lying on a different side). The lack of generalization of the RF performance across the three HAR tasks indicates the need for PA datasets with more activities and activity variations, such as this one, so researchers can improve the methods and models used for HAR. Such datasets would allow for novel algorithms to reliably differentiate small changes in activities (i.e., a reliable, robust HAR model should be able to perform well on the 42 activities in the PAAWS SimFL+Lab dataset).

Further, again in line with prior work, the performance of the models trained on the SimFL+Lab data is not the same when evaluated on the SimFL+Lab and the FL data [25, 49]. When evaluated on FL data, the performance of the RFs trained to recognize only five activities decreased by a weighted F1 of approximately 0.1 for both the wrist and thigh sensor. The performance further decreased on the activity set composed of nine activities

[3]We evaluate the results of using 126 participants data in an algorithm across five different training/evaluation sets comprised of 126 participants data to account for differences in individual movements. The average of these five train/test splits is reported. Participant IDs in each split are reported in Table 10 in the Appendix.

Table 6. Benchmarking results by number of participants and sensor location for the RFs trained using the SimFL+Lab data and evaluated using the FL data. Metrics are reported as the mean ± standard deviation for leave-one-participant-out cross validation.

Sensor	N =	5 Activities (Regrouped)		9 Activities (Regrouped)	
		W F1	M F1	W F1	M F1
Left Wrist (Top)	20	0.65 ± 0.07	0.41 ± 0.05	0.52 ± 0.25	0.19 ± 0.09
	126	0.70 ± 0.08	0.44 ± 0.06	0.68 ± 0.09	0.25 ± 0.03
	244	0.70 ± 0.08	0.44 ± 0.06	0.69 ± 0.09	0.25 ± 0.03
Right Thigh	20	0.80 ± 0.07	0.57 ± 0.07	0.79 ± 0.07	0.40 ± 0.05
	126	0.83 ± 0.08	0.60 ± 0.07	0.80 ± 0.08	0.42 ± 0.05
	244	0.83 ± 0.07	0.60 ± 0.07	0.80 ± 0.07	0.43 ± 0.06

Abbreviations: W F1 – weighted F1 score and M F1 – macro F1 score.

(a decrease of ~0.3 in both weighted and macro-averaged F1). This decrease in performance indicates that the movements captured in each dataset are distinct, possibly because true free-living behavior is more complex than behavior simulated in a lab.

4.2.2 Benchmarking the PAAWS R1 FL Dataset (Activity Types). To benchmark a HAR algorithm using the activities in the PAAWS R1 FL dataset, we considered two recognition tasks: (1) simple, re-grouped activity labels and (2) complex, semi-grouped activity labels. For the simple grouped activities task, we considered the same five activities from the first SimFL+Lab dataset benchmarking task: *sitting*, *standing*, *lying down*, *walking*, and *biking*. For the complex, semi-grouped task, we considered eleven activities: *chores*, *grooming*, *cooking*, *eating/drinking*, *standing*, *sitting*, *lying down*, *walking*, *biking*, *exercising*, and *driving*. These eleven grouped activities comprised 35 of the 38 PA type labels in the FL dataset. Because there were so few samples, we removed *Kneeling_With_Movement* and *Kneeling_Still*. Additionally, we did not include instances of *Puttering_Around* because that label is used to describe rapidly changing activities comprised of both walking and standing. The mapping schemes used in both tasks can be found in Table 11 in the Appendix. We summarize the results of this experiment in Table 7. Class-wise results are reported in the Appendix (Fig. 8).

Table 7. Benchmarking results for the RF trained on the PAAWS R1 FL data for the left wrist and right thigh sensors. Metrics are reported as the mean ± the standard deviation for leave-one-participant-out cross validation.

Sensor	Simple Task: 5 acts		Complex Task: 11 acts	
	W F1	M F1	W F1	M F1
Left Wrist (Top)	0.72 ± 0.14	0.42 ± 0.05	0.51 ± 0.13	0.20 ± 0.03
Right Thigh	0.82 ± 0.12	0.60 ± 0.07	0.61 ± 0.10	0.29 ± 0.03

Abbreviations: W F1 – weighted F1 score and M F1 – macro F1 score.

Across the same five activities used when benchmarking the SimFL+Lab dataset (the simple task) using the same features, algorithms, and sensor locations, RFs trained and evaluated using only FL data perform worse than the RFs trained and evaluated using only SimFL+Lab data from the same 20 participants (Table 6 and Table 7). For the FL data, the macro-weighted F1 is 0.42 ± 0.05 and 0.60 ± 0.07 for the wrist and thigh respectively, a 0.25 drop from the same task using the same 20 participants with SimFL+Lab data. The weighted F1 score across both sensors also drops by ~0.1 when training and evaluating on FL data in comparison to the SimFL+Lab data. The

models trained and evaluated on only FL data, interestingly, perform roughly the same as models trained on SimFL+Lab data and applied to classify true FL data, further indicating the SimFL+Lab data and FL data are not equivalently complex. The relatively low classification rates we report using standard methods (1) likely results from the complexity of the task of recognizing activities of daily life in true free-living settings, (2) highlights the need for more publicly available FL HAR datasets that capture large amounts of variability in how activities are performed, and (3) presents a challenge for the HAR research community to overcome in future work.

4.2.3 Benchmarking the PAAWS R1 FL Dataset (Sleep/Wake). To validate the sleep labels obtained from PSG data collected during up to two nights per participant during the FL protocol, we considered the task of sleep/wake recognition using wrist accelerometer data. Specifically, we ran the SWaN algorithm [59] (a sleep, wake, and non-wear algorithm) on the wrist-worn FL accelerometer data for all participants in the PAAWS R1 FL study who contributed PSG sleep data. We compared the results of the sleep detection algorithm with the gold-standard sleep labels derived from the PSG data recorded at home across the 29 nights of sleep data. Both the PSG labels and the SWaN algorithm use a 30 s granularity; these windows, however, may not perfectly overlap. To ensure a fair evaluation, we segmented each prediction and true label into 1 s segments and computed the weighted and macro F1-score using these 1 s segments. The results of comparing the SWaN algorithm for sleep prediction against the PSG labels across both wrist sensor locations are summarized in Table 8. The SWaN algorithm performs roughly the same on the PAAWS R1 dataset as the original results (a reported weighted F1 score of 0.84 ± 0.60 [59]). Additional class-wise results of the SWaN algorithm on the PAAWS data are reported in the Appendix (Fig. 9).

Table 8. Benchmarking results for the SWaN algorithm [59] using the PSG sleep labels for validation. Metrics are reported as the mean and standard deviation across participants with PSG sleep data. Analysis of predictions for 29 nights from 15 of the 20 total FL participants.

Sensor	**Weighted F1**	**Macro F1**
Left Wrist (Top)	0.85 ± 0.06	0.69 ± 0.12
Right Wrist (Top)	0.84 ± 0.07	0.66 ± 0.15

5 DISCUSSION

The PAAWS data are primarily intended to help research communities interested in using sensor-based HAR for reliable detection of PA, SB, and sleep in health studies that have gathered raw accelerometer data from people in free-living settings (e.g., NHANES or UK Biobank). Due to its volume of data, participant age diversity, and multi-limb and multimodal sensing, the PAAWS R1 dataset provides researchers with new investigative opportunities, especially when used together with existing datasets. The PAAWS data will allow head-to-head algorithm comparisons and benchmarks—a strategy that has worked effectively to spur algorithmic advances in fields such as computer vision and natural language processing—on multi-day sensor data labeled at a second-by-second resolution. Collecting, labeling, and cleaning the dataset has proven challenging, time-consuming, and expensive, thus, we have chosen to release data incrementally. As we release additional FL and sleep data from more participants, the value of the dataset will increase.[4]

5.1 Research Opportunities Using the PAAWS R1 Dataset

The PAAWS R1 dataset is a novel dataset for both the HAR and health research communities because it includes continuously labeled, multi-location, multi-day raw sensor data for waking day activities and sleep from the same multi-generational participants.

[4]We expect the full release of the dataset to the public in 2026.

Opportunities for using the SimFL+Lab data: The diversity of the PAAWS SimFL+Lab dataset—the range of activities represented, the variability within each activity type, and the age demographics of participants—combined with the relative cleanliness of the data (due to supervised data collection) from a large selection of sensor locations provides researchers with a reliable foundation for training and comparing HAR models on lab and/or semi-naturalistic data. Ideally, a structured and systematic comparison of these models will lead to more effective approaches for real-world, free-living HAR applications that use either a sensor at a single location or combinations of sensors. The large number of sensors also allows for systematic explorations of problems such as the impact of sensor displacement (e.g., rotations around the wrist or ankle) on algorithm performance. The controlled nature of the activities simplifies the training of algorithms on important activities (e.g., various walking speeds) and postures. Beyond the dataset itself, our protocol for collecting more realistic and complex data in a supervised setting can serve as a model for future studies.

Opportunities for using the FL data: The PAAWS FL data can be used to generate and validate models to investigate existing large-scale, wrist-acceleration-only datasets, which is critical for the health community given the large datasets that have already been collected in studies such as NHANES and UK Biobank. Because this new dataset provides multi-location sensor data, however, researchers may also investigate the impacts of sensor location and multi-sensor-location algorithms on FL HAR. Laboratory based testing of HAR algorithms suggests

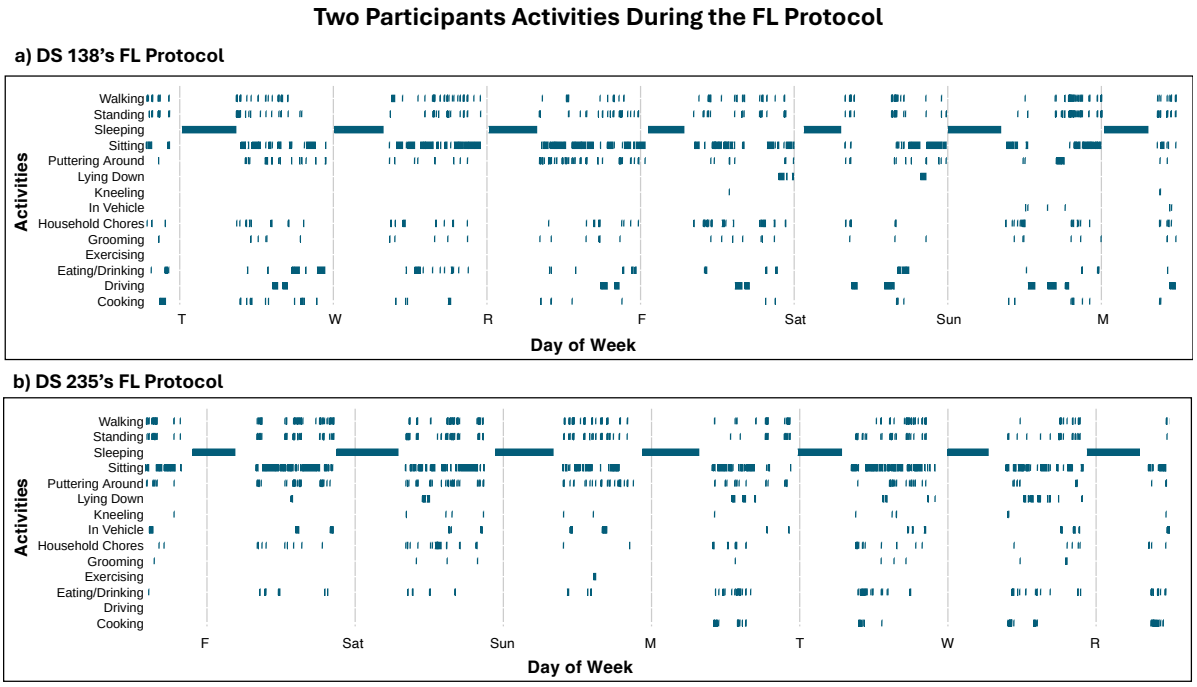

Fig. 4. The annotations of two different participants' FL data. Sleep times are from self-report labels. Top: Participant DS_138. Bottom: Participant DS_235. While both participants have similar common activities (sitting, standing, and walking), their daily activities differ (e.g., DS_138 drove substantially more). Because multiple days of ground truth data are labeled, behavioral patterns can emerge (e.g., DS_235 appears to commute frequently by public transit, ride share, or carpool). Activity groupings are listed in the Appendix (Table 12). **Abbreviations**: M - Monday, T - Tuesday, W - Wednesday, R - Thursday, F - Friday, Sat - Saturday, Sun – Sunday.

that sensor location and number of sensors impacts HAR accuracy, but datasets such as PAAWS are required to extend this work to FL settings. Further, the ***PAAWS dataset is the only continuously labeled dataset comprised of multi-sensor measurements from the same individuals across multiple days***. This provides researchers with opportunities to investigate training and evaluating HAR models within a participant across different days, or the impacts of sensor displacement on HAR algorithms over long-term FL wear. Additionally, because activities vary greatly across participants' FL behaviors, but most participants are somewhat habitual throughout their own week (Fig. 4), there may be opportunities to learn individuals' habits across multiple days to better optimize prompting for just-in-time interventions or other applications that use HAR. Further, much of the work with HAR on laboratory datasets ignores activity transitions because activities are cleanly segmented. Using the second-by-second annotation of the PAAWS R1 FL data, which captures short activities (50% of the labeled activity bouts in the dataset are under 34 s), researchers can pursue building precise HAR segmentation algorithms that appropriately detect FL activity transitions. Additionally, because PAAWS R1 includes both SimFL+Lab and FL data, researchers can use the dataset to evaluate if, and how well, their algorithms trained on data from a lab setting generalize to FL HAR. They can also explore training on different, preexisting, datasets and testing on the (more controlled) SimFL+Lab and the (less controlled) FL. Such research is critical to advance the field of activity recognition.

Opportunities for using the gold-standard sleep labels and PSG data in the FL dataset: During the FL protocol, some participants wore PSG sensors during the night—in addition to the limb motion sensors worn throughout the FL protocol—to obtain at-home PSG sleep data and accurate sleep labels, labeled at a 30 s granularity. Because recording PSG sleep data at home is known to result in longer sleep duration, better sleep efficiency, and less sleep latency than when a participant sleeps in a sleep lab [15] while providing the same PSG data quality [14, 68], the PSG sleep data can possibly be used by sleep researchers to investigate if during-day activity before a night of sleep impacts the nature of the sleep data in a way that changes a sleep detection algorithm's performance. Outside of using the raw PSG sleep data, the gold-standard sleep labels derived from the PSG data can be used to train and validate better sleep/wake detection algorithms. In addition, true sleep data (obtained from the PSG labels) might be used to train or validate algorithms intended to detect sleep during the day, such as naps (that happen occasionally in the PAAWS FL data). Sleep/wake, sleep, and sleep stage algorithms trained on other datasets can be run on the 24/7 FL data and validated against the nights with PSG data and checked for false-positive sleep during sedentary periods of the day. Further, the multi-limb accelerometer sensing recorded during the PSG sleep measurements allows for studies aimed at improving the performance of sleep stage and posture detection from raw-accelerometer data from a single sensor, non-wrist sensors, or multiple sensors.

5.2 Limitations: Sensor Type, Placement, and Wear Locations

During each protocol, participants wore research-grade accelerometers in lieu of consumer-grade multimodal wearables. The choice to use research-grade accelerometers was driven by multiple factors: (1) their batteries can last the length of the FL protocol without requiring any end-user charging, (2) they have a reliable sampling rate, (3) they are small enough to be placed on multiple limbs, and (4) they have already been used in influential health surveillance studies involving data collection from tens of thousands of participants. Labeled raw accelerometer data are essential to understanding large health datasets such as NHANES [1] and UK Biobank [24] that only include accelerometer data. As such, the PAAWS FL dataset does not include data such as heart rate data, IMU, audio, and skin conductance data that some consumer wearables can collect when charged daily. In the future, when it is more practical to collect such data 24/7 without burdening participants, more datasets may include them. For now, there is still a pressing need to develop and validate algorithms that perform HAR using only

accelerometer data, especially to allow better analysis and use of existing, valuable, unlabeled accelerometer-only datasets.

To balance participants' comfort, study compliance, and quality of data collected for the use of building better HAR models, the FL protocol used only five on-body sensor locations plus the participant's phone. (The SimFL+Lab protocol, however, used an additional fifteen sensor wear locations, including locations that may provide information on the impacts of sensor rotation on a limb.) During the FL protocol, the only sensor that remained on throughout the entirety of the protocol was the thigh sensor because it was attached to the skin with waterproof tape. This sensor alone can provide valuable information about leg orientation (and thus reliable sit/stand recognition) and movement when other sensors are off the body. Despite the limitations of sensor placement and type used in the PAAWS data collection procedures, the multi-location accelerometer data from both the SimFL+Lab and FL protocol as well as the at-home PSG sleep data will provide researchers with novel data to compare and develop HAR algorithms for PA, SB, and sleep.

5.3 Data Collection Challenges

The PAAWS dataset is one of the largest labeled PA, SB, and sleep datasets; the first release includes data from 252 participants resulting in 252 SimFL+Lab datasets and 20 FL datasets. The full dataset will include a total of 237 FL datasets (217 additional FL datasets). Collecting a dataset of this size and caliber came with many challenges such as recruiting multi-generational participants willing to wear front-facing cameras and labeling ~20k hours of video footage in the full dataset.

5.3.1 Recruiting a Large, Diverse Sample Size. We recruited 254 participants to the PAAWS study: 252 participants contributed data from the SimFL+Lab protocol, 237 participants contributed data from the FL protocol. A total of 158 participants contributed PSG sleep data during the FL protocol. To ensure this dataset was accurately labeled, each participant was asked to wear a chest-mounted, front-facing camera for the entirety of the study. The requirement to wear a front-facing, ego-centric camera during waking hours for a week is the primary reason potential participants declined participation in our study. Many potential participants were concerned about their privacy when wearing the camera or their ability to wear the camera throughout the week (i.e., they were concerned about using the camera in their workplace). Thus, the requirement to wear a camera may have introduced bias into participant recruitment. We deliberately recruited participants aged 18-80, roughly balanced across gender lines, from a variety of ethnic backgrounds. Ensuring we collected data from this diverse participant body meant extending data collection timelines and rejecting some potential participants.

5.3.2 Sensor and Camera Wear Adherence. During both data collection protocols, we faced challenges with sensor and camera wear adherence. In both protocols, participants were asked to put some sensors on their body themselves, after instruction. Nevertheless, this protocol led to several sensor orientation errors (e.g., participants put the sensors on upside down under a pant leg) or bad sensor connections (e.g., the heart rate monitor strap was in the wrong location). Additionally, during the SimFL+Lab protocol, most of the sensors were placed on participants either using clips or a tape. Occasionally, a sensor fell off at some point during the data collection protocol (e.g., a participant's waist band was loose so a clip holding a waist sensor fell off). In our protocol, each time sensors fell off, they were replaced before beginning the next activity. Sensor wear issues that occurred during the SimFL+Lab protocol are documented in the participant's notes and released with the dataset. Further, during the FL protocol, participants were instructed to wear the sensors as much as they could and told to specifically remove every sensor other than the thigh sensor during water-based activities. During daily recalls and check-ins, however, we found that several participants removed sensors outside of these specified times. Although participants were supposed to record and report all such instances, when analyzing the data we found additional instances when participants clearly removed sensors during the FL protocol. Participants

likely removed sensors for their own comfort or because they were entering a location where it would be seen as socially unacceptable to be wearing such devices. All instances of sensor non-wear that were reported or that we clearly observed when checking the data are included in the notes for each participant.

Throughout data collection, we also found variations in camera wear compliance. Missing footage resulted from two situations (1) the participant did not wear the camera or blocked the camera for privacy reasons, or (2) the participant was getting in/out of bed. Participants were asked to report any times they removed the camera to a research assistant during a daily recall session. They were also asked to self-report what they were doing during those times. These self-report activity recall data are included in participant notes in the dataset to supplement annotations of $PA_Type_Video_Unavailable/Indecipherable$. Participants also self-reported their time in bed and time asleep. In the FL study, on nights without PSG data, these self-report labels may be used to assist in training/evaluating sleep/wake algorithms, but all self-report data can have inaccurate start/stop timestamps.

5.3.3 Labeling ~20k Hours of Video Footage . The SimFL+Lab and FL protocols for all 254 participants enrolled in the PAAWS study resulted in ~20k hours of video footage to be annotated. Because we are labeling data at a second-by-second granularity, this task is complex, expensive, and labor intensive. To mitigate some of the labor strain required during annotation, we originally intended to use time lapse still-image labeling, similar to CAPTURE-24 [17], but we found that only with the video could we reliably label quick transitions that often occur in free-living behaviors as people go about their lives. Additionally, oftentimes we could only tell what someone was doing from the video footage, which allows for better determination of body motion and context. From the video footage, the most skilled annotators can annotate at a rate of approximately three hours of video footage labeled per one hour of annotation work, and many annotators are less efficient. Ideally, we would like to have two annotators annotate every video, however, even a single pass annotating the entirety of the entire PAAWS dataset will require a total of ~7.9k labor hours of annotation time (excluding training and onboarding time) [60].

6 PAAWS DATASET RELEASE TIMELINE

This paper describes the PAAWS R1 release. We have finished collecting data for the full dataset and are actively annotating, processing, and cleaning these data for future releases using the methods described in this work. Once complete, the PAAWS dataset will contain data from 237 participants completing the FL protocol (158 of who have completed at least one night of the accurately labeled sleep and PSG data). As of July 31st, 2025, we have annotated the footage of ~140 participants' FL protocols. Additionally, at this time, all the PSG sleep data has been annotated and is being processed for release. We expect the full PAAWS dataset to be cleaned and released to the public in 2026. More information about future releases and the dataset can be found at github.com/mHealth-Research-Group/paaws-benchmarking.

7 CONCLUSION

The PAAWS R1 dataset is comprised of second-by-second annotated raw sensor data from ~4 h of simulated free-living and laboratory activities (n = 252) and ~7 d of true free-living movements (n = 20), with up to two nights of accurately labeled sleep and PSG data recorded during the week-long, free-living period. The PAAWS dataset has been collected to help accelerate innovation and collaboration among those working on the next generation of wearable sensing activity recognition algorithms, especially those targeting solutions for physical activity, sedentary behavior, and sleep measurement for health research and interventions. The PAAWS R1 data is the initial release of the PAAWS dataset that will eventually include additional releases of labeled multi-limb, multi-sensor measurements for sedentary behaviors and physical activities for one full week, 24 hours per day, from 237 adults, 158 of who have at least one night of accurately labeled sleep from PSG sleep data collected at home. We have finished data collection, and we expect the entire labeled dataset will be fully released in 2026.

DATA AVAILABILITY

The PAAWS R1 dataset is available for download at https://hdl.handle.net/2047/D20806901.

CODE AVAILABILITY

To provide further transparency and encourage others to use our dataset, we release two codebooks (one for data users and one detailing more information about our data collection protocol) as well as the code, raw results, and trained model weights from our benchmarking experiments at github.com/mHealth-Research-Group/paaws-benchmarking.

ACKNOWLEDGMENTS

We would like to thank our annotators, co-op students, and participants, without who this research would not be possible. This research was supported in part by the Khoury College of Computer Sciences, Northeastern University and funded in part by the National Cancer Institute of the National Institutes of Health, grant number R01CA252966. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] National Health and Nutrition Examination Survey (NHANES). http://www.cdc.gov/nchs/nhanes, 2018.

[2] Matthew N. Ahmadi, Alok Chowdhury, Toby Pavey, and Stewart G. Trost. Laboratory-Based and Free-Living Algorithms for Energy Expenditure Estimation in Preschool Children: A Free-Living Evaluation. *PLOS ONE*, 15(5):e0233229, 2020.

[3] Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative Study on Classifying Human Activities with Miniature Inertial and Magnetic Sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

[4] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. In *The European Symposium on Artificial Neural Networks*, 2013.

[5] Luca Arrotta, Gabriele Civitarese, Riccardo Presotto, and Claudio Bettini. DOMINO: A Dataset for Context-Aware Human Activity Recognition using Mobile Devices. In *2023 24th IEEE International Conference on Mobile Data Management (MDM)*, pages 346–351, 2023.

[6] Oresti Baños, Miguel Damas, Héctor Pomares, Ignacio Rojas, Máté Attila Tóth, and Oliver Amft. A Benchmark Dataset to Evaluate Sensor Displacement in Activity Recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 1026–1035, 2012.

[7] Oresti Baños, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mHealthDroid: A Novel Framework for Agile Development of Mobile Health Applications. pages 91–98. Springer International Publishing, 2014.

[8] Ling Bao and Stephen S. Intille. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive Computing*, pages 1–17, 2004.

[9] Eugen Berlin and Kristof Van Laerhoven. Detecting Leisure Activities with Dense Motif Discovery. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 250–259, 2012.

[10] Richard B. Berry, Stuart F. Quan, Alexandre R. Abreu, Marietta L. Bibbs, Lourdes DelRosso, Susan M. Harding, Molly-Min Mao, David T. Plante, Mark R. Pressman, Matthew M. Troester, and Bradley V. Vaughn. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. *The American Academy of Sleep Medicine*, 2.6, 2020.

[11] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(4), 2024.

[12] Gino Brunner, Darya Melnyk, Birkir Sigfússon, and Roger Wattenhofer. Swimming Style Recognition and Lap Counting Using a Smartwatch and Deep Learning. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 23–31, 2019.

[13] Barbara Bruno, Fulvio Mastrogiovanni, Antonio Sgorbissa, Tullio Vernazza, and Renato Zaccaria. Analysis of Human Behavior Recognition Algorithms Based on Acceleration Data. In *IEEE International Conference on Robotics and Automation*, pages 1602–1607, 2013.

[14] Marie Bruyneel, Walter Libert, Lieveke Ameye, and Vincent Ninane. Comparison Between Home and Hospital Set-Up for Unattended Home-Based Polysomnography: A Prospective Randomized Study. *Sleep Medicine*, 16(11):1434–1438, 2015.

[15] Marie Bruyneel, Christina Sanida, Geneviève Art, Walter Libert, Laurent Cuvelier, Marianne Paesmans, Roger Sergysels, and Vincent Ninane. Sleep Efficiency During Sleep Studies: Results of a Prospective Study Comparing Home-Based and In-Hospital Polysomnography. *Journal of Sleep Research*, 20(1.2):201–206, 2011.

[16] Marzia Cescon, Divya Choudhary, Jordan E. Pinsker, Vikash Dadlani, Mei Mei Church, Yogish C. Kudva, Francis J. Doyle Iii, and Eyal Dassau. Activity Detection and Classification from Wristband Accelerometer Data Collected on People with Type 1 Diabetes in Free-Living Conditions. *Computers in Biology and Medicine*, 135, 2021.

[17] Shing Chan, Yuan Hang, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. CAPTURE-24: A Large Dataset of Wrist-Worn Activity Tracker Data Collected in the Wild for Human Activity Recognition. *Scientific Data*, 11(1), 2024.

[18] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. The OPPORTUNITY Challenge: A Benchmark Database for On-Body Sensor-Based Activity Recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

[19] Chen Chen, Roozbe Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. In *IEEE International Conference on Image Processing (ICIP)*, pages 168–172, 2015.

[20] Zhengming Chen, Junshi Chen, Rory Collins, Yu Guo, Richard Peto, Fan Wu, Liming Li, and on behalf of the China Kadoorie Biobank collaborative group. China Kadoorie Biobank of 0.5 Million People: Survey Methods, Baseline Characteristics and Long-Term Follow-Up. *International Journal of Epidemiology*, 40(6):1652–1666, 2011.

[21] Mathias Ciliberto, Vitor Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. OPPORTUNITY++: A Multimodal Dataset for Video- and Wearable, Object and Ambient Sensors-Based Human Activity Recognition. *Frontiers in Computer Science*, 3, 2021.

[22] Scott E. Crouter, James R. Churilla, and David R. Bassett. Estimating Energy Expenditure Using Accelerometers. *Eur. J. Appl. Physiol.*, 98(6):601–612, 2006.

[23] Scott E. Crouter, Diane M. DellaValle, Jere D. Haas, Edward A. Frongillo, and David R. Bassett. Validity of ActiGraph 2-Regression Model, Matthews Cut-Points, and NHANES Cut-Points for Assessing Free-Living Physical Activity. *Journal of Physical Activity and Health*, 10(4):504–14, 2013.

[24] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H. Granat, Tom White, Vincent T. Van Hees, Michael I. Trenell, Christoper G. Owen, Stephen J. Preece, Rob Gillions, Simon Sheard, Tim Peakman, Soren Brage, and Nicholas J. Wareham. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE*, 12(2), 2017.

[25] Arindam Dutta, Owen Ma, Meynard Toledo, Alberto F. Pregonero, Barbara E. Ainsworth, Matthew P. Buman, and Daniel W. Bliss. Identifying Free-Living Physical Activities Using Lab-Based Models with Wearable Accelerometers. *Sensors*, 18(11), 2018.

[26] Alexander Hoelzemann, Marius Bock, Ericka Andrea Valladares Bastías, Salma El Ouazzani Touhami, Kenza Nassiri, and Kristof Van Laerhoven. A Data-Driven Study on the Hawthorne Effect in Sensor-Based Human Activity Recognition. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, UbiComp/ISWC '23 Adjunct, page 486–491, 2023.

[27] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of Activity Patterns Using Topic Models. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 10–19, 2008.

[28] Espen A. F. Ihlen, Aner Weiss, Jorunn L. Helbostad, and Jeffrey M. Hausdorff. The Discriminant Value of Phase-Dependent Local Dynamic Stability of Daily Life Walking in Older Adult Community-Dwelling Fallers and Nonfallers. *BioMed Research International*, 2015:1–11, 2015.

[29] Kristen L. Knutson, Armand M. Ryden, Bryce A. Mander, and Eve Van Cauter. Role of Sleep Duration and Quality in the Risk and Severity of Type 2 Diabetes Mellitus. *Archives of Internal Medicine*, 166(16):1768–1774, 2006.

[30] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity Recognition Using Cell Phone Accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2010.

[31] Fernando De la Torre Frade, Jessica K. Hodgins, Adam W. Bargteil, Xavier Martin Artal, Justin C. Macey, Alexandre Collado I Castells, and Josep Beltran. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. Technical Report CMU-RI-TR-08-22, 2008.

[32] I-Min Lee, Eric J. Shiroma, Felipe Lobelo, Pekka Puska, Steven N. Blair, and Peter T. Katzmarzyk. Effect of Physical Inactivity on Major Non-Communicable Diseases Worldwide: An Analysis of Burden of Disease and Life Expectancy. *The Lancet*, 380(9838):219–229, 2012.

[33] Heike Leutheuser, Dominik Schuldhaus, and Bjoern M. Eskofier. Hierarchical, Multi-Sensor Based Classification of Daily Life Activities: Comparison with State-of-the-Art Algorithms Using a Benchmark Dataset. *PLOS ONE*, 8(10), 2013.

[34] Bin Luo, Qi Qiu, Tao Zhu, and Zhenyu Liu. CAPPIMU: A Composite Activities Dataset for Human Activity Recognition Utilizing Plantar Pressure and IMU Sensors. In *Intelligent Information Processing XII*, pages 87–100. Springer Nature Switzerland, 2024.

[35] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting Sensory Data against Sensitive Inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, W-P2DS'18, 2018.

[36] Adria Mallol-Ragolta, Anastasia Semertzidou, Maria Pateraki, and Björn Schuller. harAGE: A Novel Multimodal Smartwatch-based Dataset for Human Activity Recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–07, 2021.

[37] Andrea Mannini, Stephen S. Intille, Mary Rosenberger, Angelo M. Sabatini, and William Haskell. Activity Recognition Using a Single Accelerometer Placed at the Wrist or Ankle. *Medicine Science in Sports Exercise*, 45(11):2193–2203, 2013.

[38] J. Margarito, R. Helaoui, A. M. Bianchi, F. Sartor, and A. G. Bonomi. User-Independent Recognition of Sports Activities From a Single Wrist-Worn Accelerometer: A Template-Matching-Based Approach. *IEEE Transactions on Biomedical Engineering*, 63(4):788–796, 2016.

[39] R. Mattfeld, E. Jesch, and A. Hoover. A New Dataset for Evaluating Pedometer Performance. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 865–869, 2017.

[40] Gillian McLellan, Rosie Arthur, and Duncan S. Buchan. Wear Compliance, Sedentary Behaviour and Activity in Free-Living Children from Hip- and Wrist-Mounted ActiGraph GT3X+ Accelerometers. *Journal of Sports Sciences*, 36(21):2424–2430, 2018.

[41] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones. *Applied Sciences*, 7(10):1101, 2017.

[42] Adnan Nadeem, Amir Mehmood, and Kashif Rizwan. A Dataset Build Using Wearable Inertial Measurement and ECG Sensors for Activity Recognition, Fall Detection and Basic Heart Anomaly Detection System. *Data in Brief*, 27:104717, 2019.

[43] Henry Friday Nweke, Ying Wah Teh, Uzoma Rita Alo, and Ghulam Mujtaba. Analysis of Multi-Sensor Fusion for Mobile and Wearable Sensor Based Human Activity Recognition. In *Proceedings of the International Conference on Data Processing and Applications*, page 22–26, 2018.

[44] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60, 2013.

[45] Daniel Olguín Olguín and Alex Sandy Pentland. Human Activity Recognition: Accuracy Across Common Locations for Wearable Sensors. In *IEEE 10th International Symposium on Wearable Computers*, pages 11–14, 2006.

[46] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. COSMO: Contrastive Fusion Learning with Small Data for Multimodal Human Activity Recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, page 324–337, 2022.

[47] Attila Reiss and Didier Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109, 2012.

[48] Alessio Rossi, Eleonora Da Pozzo, Dario Menicagli, Chiara Tremolanti, Corrado Priami, Alina Sîrbu, David A. Clifton, Claudia Martini, and Davide Morelli. A Public Dataset of 24-h Multi-Levels Psycho-Physiological Responses in Young Healthy Adults. *Data*, 5(4), 2020.

[49] Jeffer Eidi Sasaki, Amanda M. Hickey, John W. Staudenmayer, Dinesh John, Jane A. Kent, and Patty S. Freedson. Performance of Activity Classification Algorithms in Free-Living Older Adults. *Medicine Science in Sports Exercise*, 48(5):941–950, 2016.

[50] Severin Schricker, Nico Schmid, Moritz Schanz, Martin Kimmel, and Mark D. Alscher. In-Hospital Physical Activity Measured with a New Bosch Accelerometer Sensor System (version 1.0). *PhysioNet*. RRID:SCR_007345, 2020.

[51] Ahmed Sharshar, Ahmed Fayez, Ahmed Abo Eitta, and Walid Gomaa. MM-DOS: A Novel Dataset Of Workout Activities. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[52] Muhammad Shoaib, Stephan Bosch, Ozlem Incel, Hans Scholten, and Paul Havinga. Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors. *Sensors*, 16(4), 2016.

[53] Niloy Sikder and Abdullah-Al Nahid. KU-HAR: An Open Dataset for Heterogeneous Human Activity Recognition. *Pattern Recognition Letters*, 146:46–54, 2021.

[54] Salwa O. Slim, Ayman Atia, Marwa M.A. Elfattah, and Mostafa-Sami M.Mostafa. Survey on Human Activity Recognition based on Acceleration Data. *International Journal of Advanced Computer Science and Applications*, 10(3), 2019.

[55] Marija Stojchevska, Mathias De Brouwer, Martijn Courteaux, Femke Ongenae, and Sofie Van Hoecke. From Lab to Real World: Assessing the Effectiveness of Human Activity Recognition and Optimization through Personalization. *Sensors*, 23(10), 2023.

[56] S. J. Strath, Jr. Bassett, D. R., and A. M. Swartz. Comparison of MTI Accelerometer Cut-Points for Predicting Time Spent in Physical Activity. *Int J Sports Med*, 24(4):298–303, 2003.

[57] Timo Sztyler and Heiner Stuckenschmidt. On-Body Localization of Wearable Devices: An Investigation of Position-Aware Activity Recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9, 2016.

[58] Qu Tang, Dinesh John, Binod Thapa-Chhetry, Diego J. Arguello, and Stephen Intille. Posture and Physical Activity Detection: Impact of Number of Sensors and Feature Type. *Med Sci Sports Exerc*, 52(8):1834–1845, 2020.

[59] Binod Thapa-Chhetry, Diego Jose Arguello, Dinesh John, and Stephen Intille. Detecting Sleep and Nonwear in 24-h Wrist Accelerometer Data from the National Health and Nutrition Examination Survey. *Med Sci Sports Exerc*, 54(11):1936–1946, 2022.

[60] Hoan Tran, Veronika Potter, Umberto Mazzucchelli, Dinesh John, and Stephen Intille. Towards Practical, Best Practice Video Annotation to Support Human Activity Recognition. In *Proceedings of the 9th International Workshop on Annotation of Real World Data for Artificial Intelligent Systems (ARDUOUS 2025)*, 2025.

[61] Richard P. Troiano, James J. McClain, Robert J. Brychta, and Kong Y. Chen. Evolution of Accelerometer Methods for Physical Activity Research. *British Journal of Sports Medicine*, 48(13):1019–1023, 2014.

[62] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches. *IEEE Pervasive Computing*, 16(4):62–74, 2017.

[63] George Vavoulas, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pediaditis, and Manolis Tsiknakis. The MobiFall Dataset: An Initial Evaluation of Fall Detection Algorithms Using Smartphones. In *13th IEEE International Conference on BioInformatics and BioEngineering*,

pages 1–4, 2013.
[64] Aiguo Wang, Guilin Chen, Jing Yang, Shenghui Zhao, and Chih-Yung Chang. A Comparative Study on Human Activity Recognition Using Inertial Sensors in a Smartphone. *IEEE Sensors Journal*, 16(11):4566–4578, 2016.
[65] Yan Wang, Shuang Cang, and Hongnian Yu. A Survey on Wearable Sensor Modality Centred Human Activity Recognition in Health Care. *Expert Systems with Applications*, 137:167–190, 2019.
[66] Yan Yan, Yi-Chun Huang, Jinjin Zhao, Yu-Shi Liu, Liang Ma, Jing Yang, Xu-Dong Yan, Jing Xiong, and Lei Wang. Topological Nonlinear Analysis of Dynamical Systems in Wearable Sensor-Based Human Physical Activity Inference. *IEEE Transactions on Human-Machine Systems*, 53(4):792–801, 2023.
[67] Allen Y. Yang, Roozbeh Jafari, S. Shankar Sastry, and Ruzena Bajcsy. Distributed Recognition of Human Actions Using Wearable Motion Sensor Networks. *Journal of Ambient Intelligence and Smart Environments*, 1(2):103–115, 2009.
[68] Dae Wui Yoon, Il-Hee Hong, Inkyung Baik, and Hyun-Woo Shin. Evaluation of the Feasibility and Preference of Nox-A1 Type 2 Ambulatory Device for Unattended Home Sleep Test: A Randomized Crossover Study. *Sleep and Biological Rhythms*, 17(3):297–304, 2019.
[69] Mi Zhang and Alexander A. Sawchuk. USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, page 1036–1043, 2012.

A Appendix: Additional Benchmarking Details and Results

A.1 Model Details

To benchmark the SimFL+Lab and FL datasets in the PAAWS R1 release (Sec. 4.2.1 and Sec. 4.2.2), we use a random forest model. We use the sklearn v1.0.2 RandomForestClassifier with 1000 trees and otherwise default settings. To benchmark the sleep labels obtained from the PSG data collected during the FL protocol, we use the SWaN algorithm [59] on the wrist sensors with default settings.

A.2 Feature Extraction

We extracted 16 different features across each channel of the accelerometer signal and the signal magnitude yielding a total of 59 features (Table 9). Correlation was only computed between pairs of channels in the raw accelerometer signal, not the signal magnitude.

Table 9. The 16 features extracted from 10 s non-overlapping windows of raw accelerometer data. We used these features when benchmarking the SimFL+Lab and FL datasets (Sec. 4.2.1 and 4.2.2).

Feature Domain	**#**	**Name**
Time	1	Min
	2	Max
	3	Mean
	4	Variance
	5	Standard deviation
	6	Skewness
	7	Kurtosis
	8	25th percentile
	9	Median (50th percentile)
	10	75th percentile
	11	Correlation
	12	Zero crossings
	13	Energy
Frequency	14	Dominant frequency
	15	Dominant magnitude
	16	Entropy

A.3 Data Exclusion and Participants

During training and evaluation, we excluded any participants' data that were partial (e.g., a sensor fell off during the weeklong FL protocol), missing, or had a known orientation issue.

When benchmarking the SimFL+Lab dataset, we considered training set sizes comprised of 20 participants (who also have FL datasets in the PAAWS R1 dataset), 126 participants (half of the entire dataset), and all participants (244 participants after data exclusion). To ensure no particular training set impacted the results when training and evaluating on data from 126 participants, we randomly selected five different sets of 126 participants from the included SimFL+Lab data (i.e., 126 participants from 244). For benchmarking the FL data (and evaluating models trained on only SimFL+Lab data on FL data), we used the data from 18 participants. For benchmarking the PSG sleep labels, we used data from 15 participant from whom we collected PSG sleep from during the FL protocol. The participants included in each training/evaluation set are detailed in Table 10.

Table 10. The datasets used in each training/evaluation set during the benchmarking experiments.

Benchmarking Dataset	***N* =**	**Participant IDs (DS_)**
SimFL+Lab (Sec. 4.2.1)	20	10, 36, 37, 38, 39, 42, 44, 48, 49, 51, 58, 59, 87, 138, 139, 140, 235, 239, 240, 246
	126 (set 1)	229, 258, 170, 238, 296, 266, 48, 61, 254, 15, 290, 59, 172, 57, 88, 113, 77, 268, 246, 235, 305, 160, 104, 218, 297, 187, 120, 280, 262, 51, 56, 279, 41, 173, 45, 298, 265, 105, 84, 101, 267, 278, 112, 206, 64, 94, 285, 12, 68, 301, 22, 133, 70, 152, 36, 252, 275, 149, 272, 294, 286, 241, 153, 273, 200, 190, 231, 287, 189, 93, 146, 23, 20, 293, 44, 184, 62, 185, 129, 28, 97, 119, 210, 127, 53, 199, 26, 289, 302, 10, 25, 257, 216, 242, 80, 107, 230, 32, 122, 264, 193, 208, 130, 151, 21, 128, 34, 244, 111, 109, 211, 55, 168, 42, 192, 66, 217, 271, 118, 65, 39, 74, 14, 239, 31, 19
	126 (set 2)	86, 11, 267, 109, 278, 50, 264, 289, 41, 136, 126, 15, 70, 184, 235, 199, 179, 98, 10, 202, 216, 218, 133, 304, 67, 271, 19, 256, 79, 217, 122, 275, 124, 31, 34, 120, 112, 290, 24, 293, 52, 212, 286, 138, 296, 258, 226, 69, 91, 297, 214, 40, 153, 94, 114, 175, 58, 160, 168, 215, 84, 56, 77, 111, 223, 238, 63, 147, 113, 64, 14, 49, 48, 140, 195, 68, 206, 61, 213, 62, 269, 229, 240, 46, 299, 80, 78, 21, 25, 273, 220, 51, 252, 65, 208, 12, 200, 93, 87, 210, 20, 33, 268, 139, 96, 105, 74, 174, 72, 243, 270, 295, 88, 132, 219, 211, 16, 279, 53, 148, 60, 115, 176, 18, 73, 100
	126 (set 3)	132, 74, 264, 271, 160, 42, 286, 92, 28, 269, 259, 267, 278, 111, 215, 73, 181, 174, 49, 204, 220, 66, 12, 17, 190, 293, 156, 262, 207, 99, 136, 94, 125, 153, 246, 279, 287, 161, 148, 234, 180, 43, 106, 260, 189, 200, 226, 230, 14, 146, 152, 302, 47, 277, 147, 281, 119, 22, 62, 19, 60, 67, 48, 109, 35, 112, 210, 256, 72, 218, 16, 36, 64, 129, 202, 219, 253, 229, 266, 244, 33, 53, 294, 27, 285, 83, 57, 187, 103, 176, 139, 296, 206, 183, 85, 32, 96, 257, 118, 184, 282, 238, 55, 175, 108, 15, 77, 188, 46, 248, 235, 169, 179, 275, 216, 38, 177, 39, 172, 151, 98, 252, 258, 168, 284, 289

Continued on next page

Table 10. The datasets used in each training/evaluation set during the benchmarking experiments.

Benchmarking Dataset	***N* =**	**Participant IDs (DS_)**
	126 (set 4)	77, 117, 139, 82, 273, 96, 152, 110, 219, 213, 114, 68, 188, 102, 78, 32, 258, 129, 99, 86, 216, 38, 113, 207, 246, 27, 71, 195, 17, 122, 294, 171, 223, 278, 211, 254, 277, 59, 73, 21, 115, 138, 244, 133, 36, 84, 58, 28, 11, 180, 226, 112, 123, 151, 88, 269, 23, 127, 81, 160, 106, 52, 193, 289, 282, 176, 47, 299, 266, 264, 168, 53, 231, 229, 296, 256, 97, 179, 126, 242, 80, 217, 279, 34, 297, 159, 267, 268, 10, 224, 45, 55, 265, 109, 293, 65, 202, 262, 164, 92, 305, 271, 252, 173, 25, 120, 260, 298, 287, 46, 69, 155, 303, 31, 174, 290, 154, 234, 237, 74, 149, 57, 128, 62, 181, 169
	126 (set 5)	183, 89, 173, 85, 16, 13, 202, 270, 120, 169, 15, 276, 127, 124, 140, 61, 188, 164, 244, 111, 231, 10, 64, 68, 298, 172, 283, 109, 246, 146, 171, 36, 267, 60, 190, 154, 210, 67, 156, 223, 49, 37, 52, 208, 206, 295, 43, 83, 287, 212, 195, 130, 98, 277, 65, 57, 22, 101, 26, 263, 152, 248, 80, 291, 39, 280, 126, 161, 239, 105, 112, 296, 56, 62, 14, 81, 304, 35, 129, 100, 92, 285, 168, 72, 32, 84, 20, 265, 63, 31, 106, 179, 115, 34, 21, 28, 79, 235, 230, 18, 260, 139, 305, 102, 82, 290, 200, 138, 38, 242, 297, 122, 87, 301, 207, 180, 90, 104, 147, 176, 229, 204, 181, 219, 160, 119
	244	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 117, 118, 119, 120, 122, 123, 124, 125, 126, 127, 128, 129, 130, 132, 133, 136, 138, 139, 140, 146, 147, 148, 149, 151, 152, 153, 154, 155, 156, 159, 160, 161, 164, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 179, 180, 181, 183, 184, 185, 187, 188, 189, 190, 192, 193, 195, 198, 199, 200, 201, 202, 204, 206, 207, 208, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 223, 224, 226, 229, 230, 231, 234, 235, 237, 238, 239, 240, 241, 242, 243, 244, 246, 248, 252, 253, 254, 256, 257, 258, 259, 260, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 289, 290, 291, 293, 294, 295, 296, 297, 298, 299, 301, 302, 303, 304, 305
FL (Sec. 4.2.2)	18	10, 36, 37, 38, 39, 42, 44, 48, 49, 51, 58, 59, 87, 138, 139, 140, 240, 246
PSG Labels (Sec. 4.2.3)	15	10, 36, 37, 38, 39, 42, 48, 49, 87, 138, 139, 235, 239, 240, 246

A.4 Activity Mapping Schemes

Before training and evaluating our models to benchmark the PAAWS R1 dataset, we developed re-grouped activity classes. These mapping schemes are defined in Table 11.

Table 11. The mapping schemes used to re-group the raw PAAWS labels into activity classes for the benchmarking in Fig. 4. Unless otherwise noted, the raw PAAWS activity label is a PA Type label.

Re-grouped Activity Label	**Raw PAAWS Activity Label**
Biking	*Stationary_Biking_300_Lab*, *Cycling_Active_Pedaling_Regular_Bicycle*, *Exercising_Gym_Stationary_Bicycle*, *Cycling_Active_Pedaling_Stationary_Bicycle*
Cooking	*Chopping_Food_Lab*, HLB: *Cooking/Prepping_Food*
Driving	HLB: *In_Transit_Driving_Car*
Eating/Drinking	HLB: *Eating/Dining*
Exercising (Gym Exercises)	*Playing_Frisbee*, HLB: *Playing_Exergame*, HLB: *Playing_Sports/Games*, *Doing_Resistance_Training_Free_Weights*, *Exercising_Gym_Other*, *Doing_Resistance_Training_Other*, *Exercising_Gym_Treadmill*, *Arm_Curls_Lab*, *Doing_Martial_Arts*, *Push_Up_Modified_Lab*, *Doing_Resistance_Training*, *Push_Up_Lab*, *Ab_Crunches_Lab*, *Machine_Leg_Press_Lab*, *Machine_Chest_Press_Lab*, *Running_Non_Treadmill*, *Running_Treadmill*
Grooming	*Washing_Hands*, *Brushing_Teeth*, *Brushing/Combing/Tying_Hair*, *Washing_Face*, *Applying_Makeup*, *Flossing_Teeth*, *Blowdrying_Hair*
Household Chores	*Dry_Mopping*, HLB: *Cleaning*, *Dusting*, *Doing_Common_Housework_Light*, *Folding_Clothes*, *Doing_Common_Housework*, *Ironing*, *Doing_Dishes*, *Loading/Unloading_Washing_Machine/Dryer*, *Doing_Home_Repair_Light*, *Organizing_Shelf/Closet*, *Doing_Home_Repair*, *Putting_Clothes_Away*, *Packing/Unpacking_Something*, *Sweeping*, *Vacuuming*, *Stand_Shelf_Load_Lab*, *Stand_Shelf_Unload_Lab*, *Washing_Dishes_Lab*
In Vehicle	HLB: *In_Transit_Passive_Car*, *In_Transit_Passive_Train/Bus/Plane*
Kneeling	*Kneeling_Still*, *Kneeling_With_Movement*
Lying Down	*Lying_Still*, *Lying_On_Left_Side_Lab*, *Lying_On_Right_Side_Lab*, *Lying_On_Back_Lab*, *Lying_On_Stomach_Lab*, *Lying_With_Movement*
Puttering Around	*Puttering_Around*
Sitting	*Sitting_Still*, *Sit_Typing_Lab*, *Sit_Recline_Web_Browse_Lab*, *Sit_Writing_Lab*, *Sit_Recline_Talk_Lab*, *Sitting_With_Movement*
Standing	*Standing_Still*, *Stand_Conversation_Lab*, *Standing_With_Movement*
Walking	*Walking*, *Walking_Slow*, *Walking_Fast*, *Walking_Up_Stairs*, *Walking_Down_Stairs*, *Walking_Treadmill*, *Treadmill_2mph_Lab*, *Treadmill_3mph_Conversation_Lab*, *Treadmill_3mph_Drink_Lab*, *Treadmill_3mph_Free_Walk_Lab*, *Treadmill_3mph_Briefcase_Lab*, *Treadmill_3mph_Hands_Pockets_Lab*, *Treadmill_4mph_Lab*

A.5 Class Wise Analysis (Confusion Matrices)

Aggregated (across all participants) confusion matrices for each experiment run to benchmark the SimFL+Lab and FL sensor data as well as the PSG sleep labels are reported in Figures 5, 6, 8, and 9.

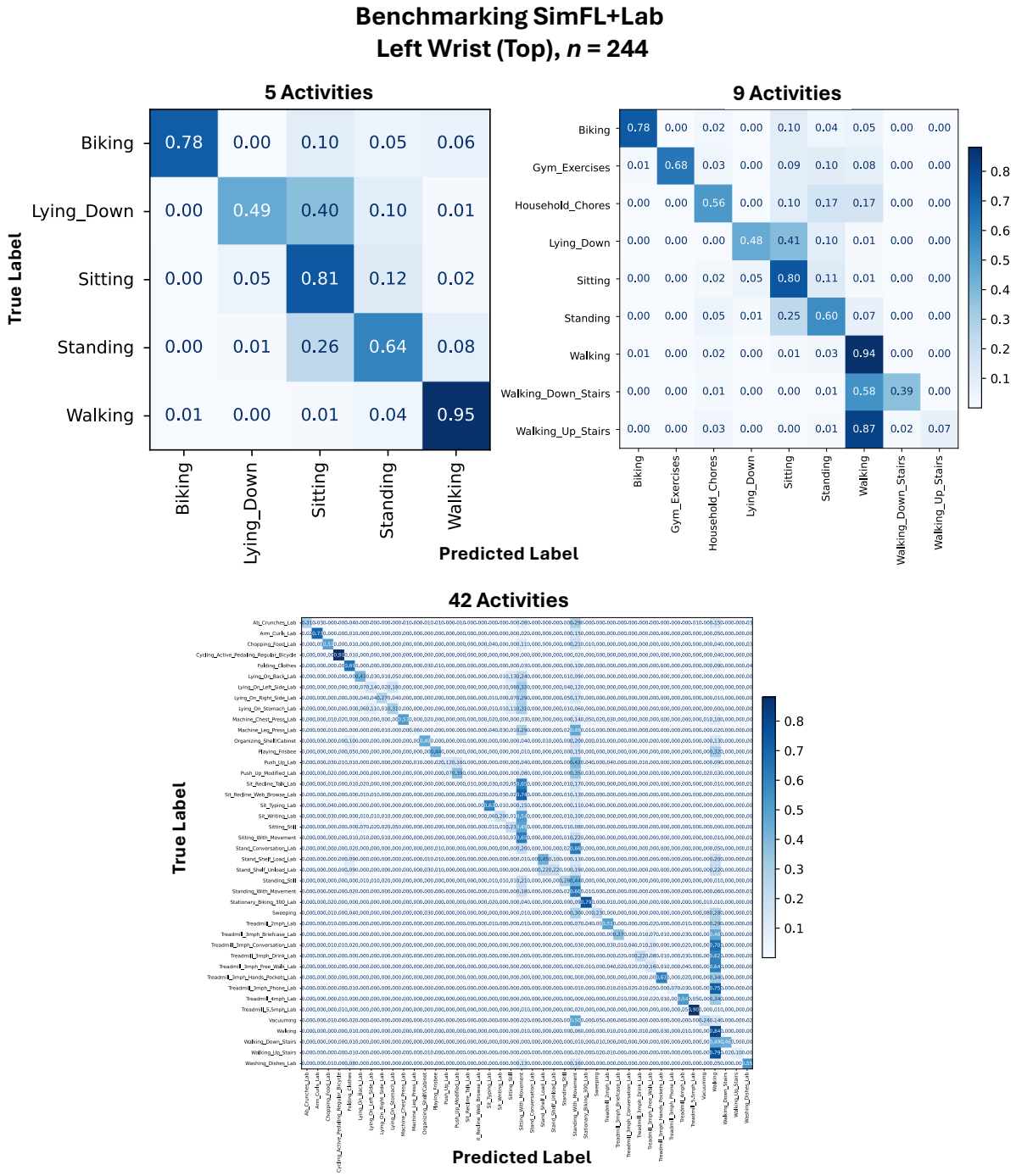

Fig. 5. The aggregate confusion matrices for benchmarking the PAAWS SimFL+Lab accelerometer data using a RF by number of activities for the training set from 244 participants (entire SimFL+Lab data) for the left wrist (top) sensor.

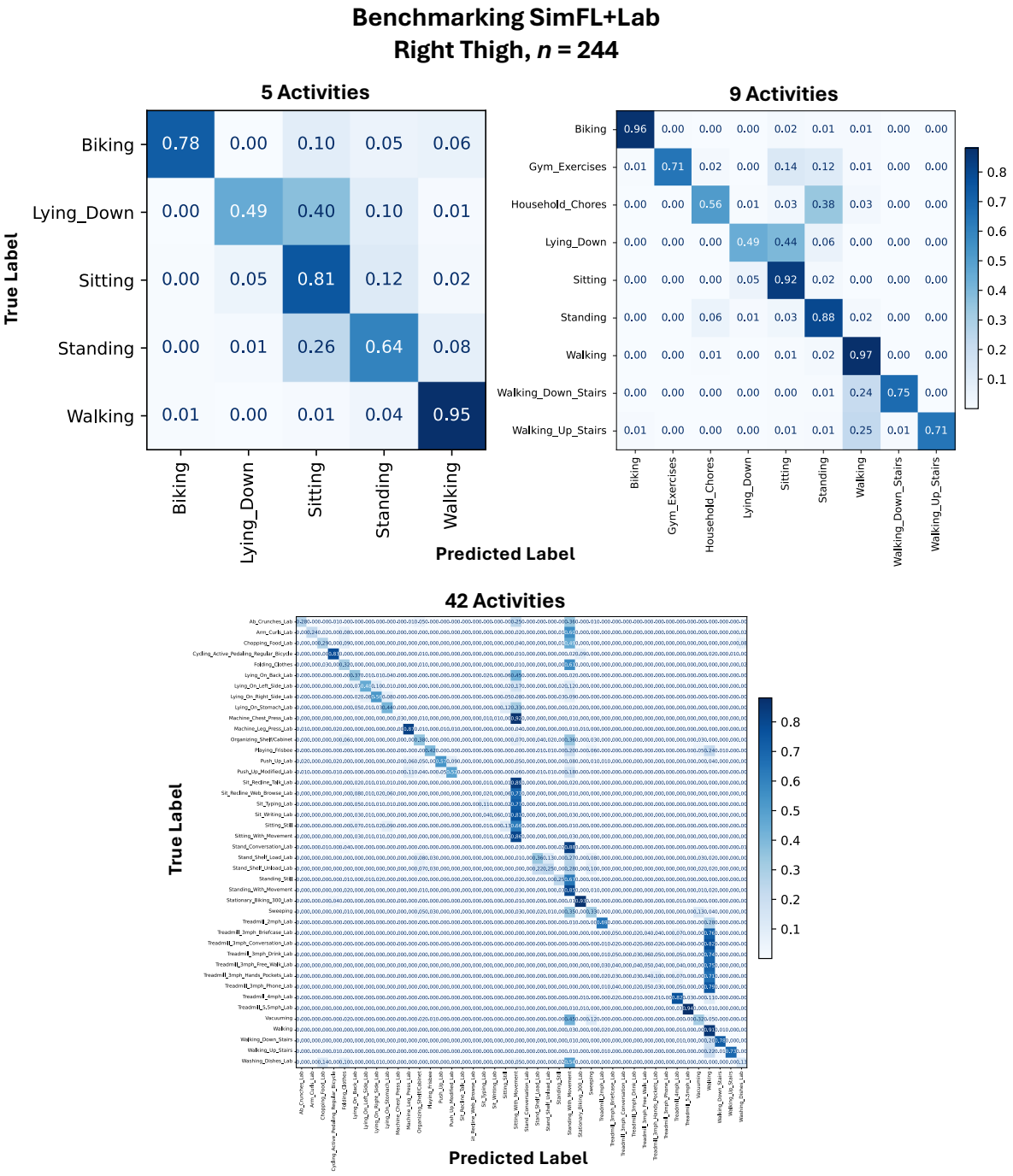

Fig. 6. The aggregate confusion matrices for benchmarking the PAAWS SimFL+Lab accelerometer data using a RF by number of activities for the training set from 244 participants (entire SimFL+Lab data) for the right thigh sensor.

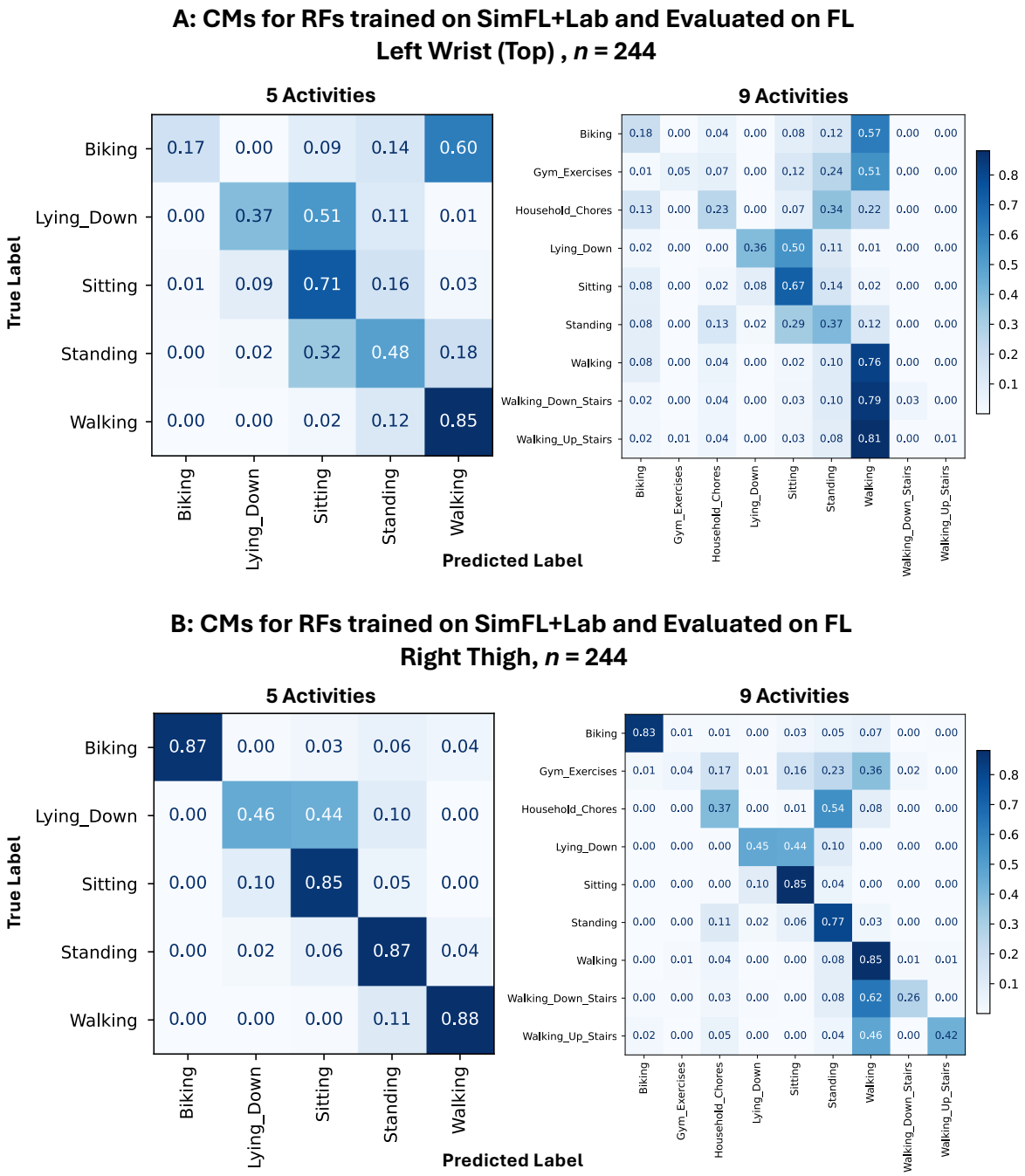

Fig. 7. The aggregate confusion matrices for benchmarking the PAAWS SimFL+Lab and FL accelerometer data using a RF by sensor location and number of activities for the models trained with 244 participants' data (the entire dataset).

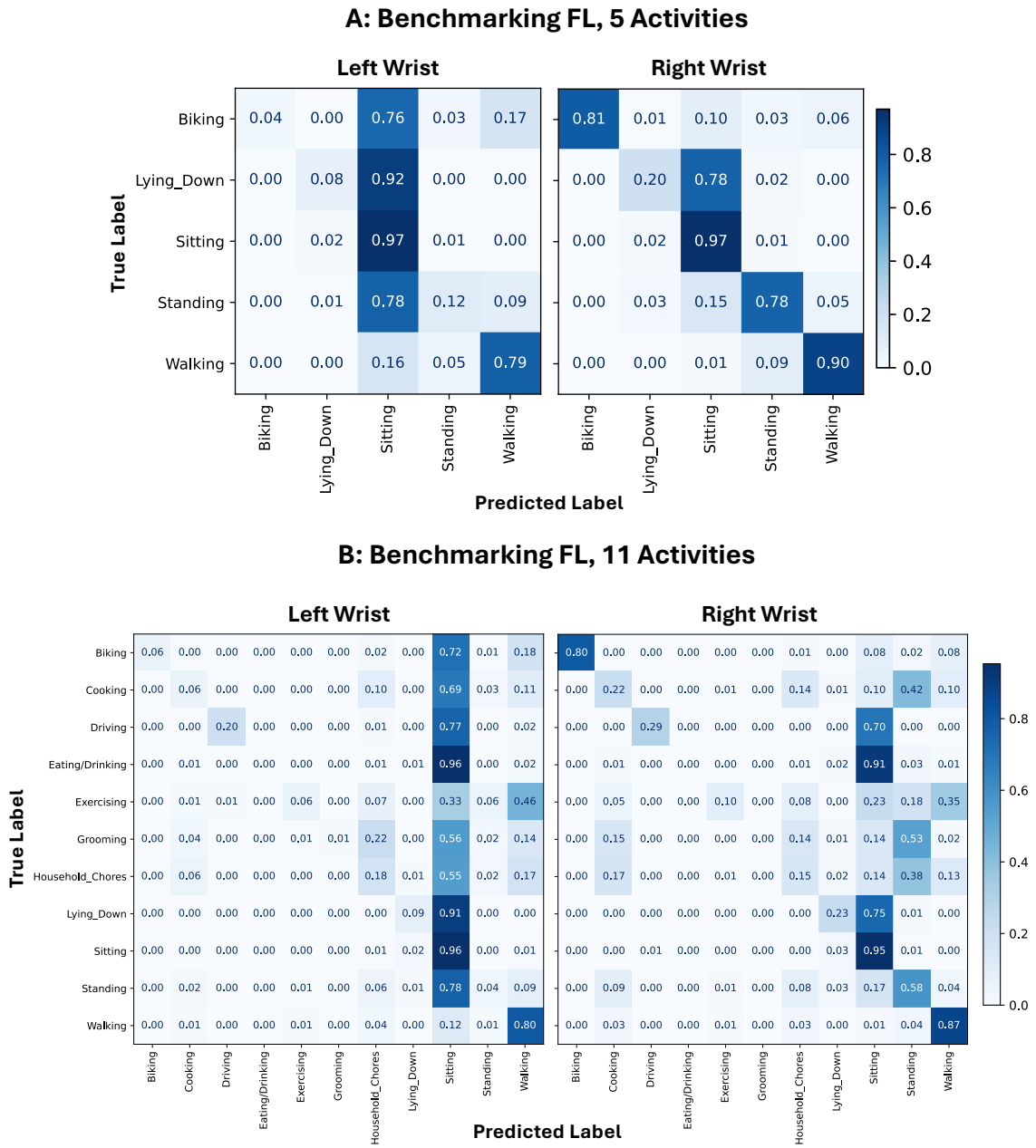

Fig. 8. The aggregate confusion matrices for benchmarking the PAAWS R1 FL accelerometer data using a RF by sensor location and number of activities.

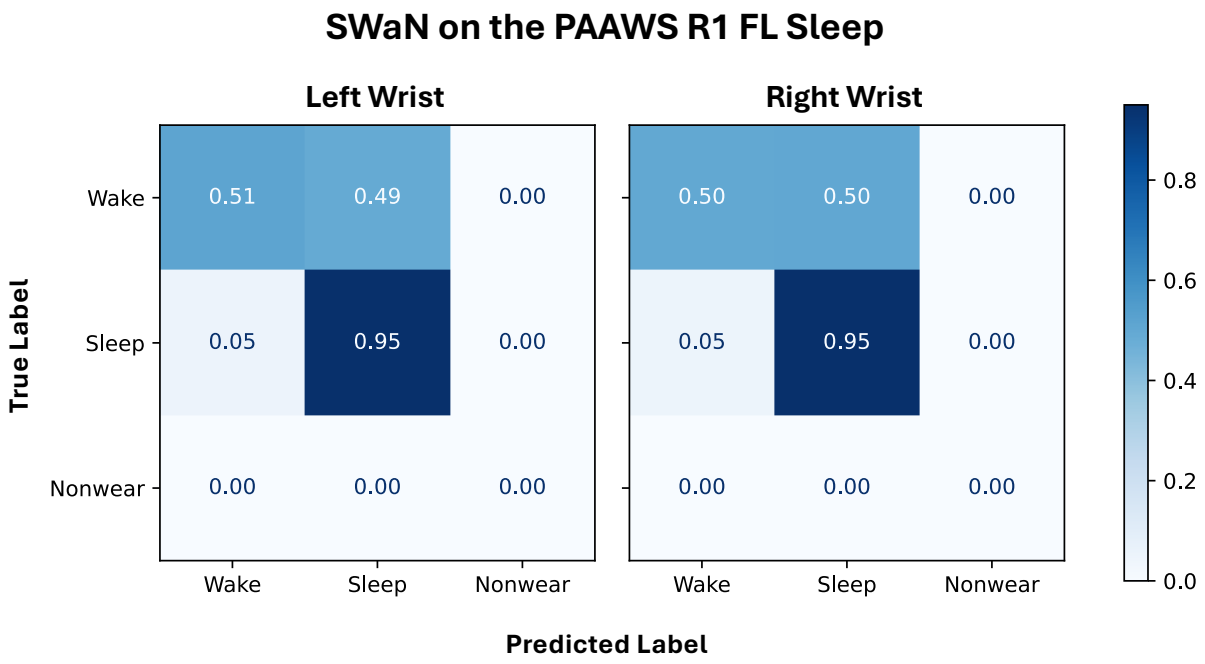

Fig. 9. The aggregate confusion matrices for benchmarking the PSG sleep labels using the SWaN algorithm by sensor locations.